BOSTON NAMING TEST WITH LATENCIES (BNT-L)

Margaret Anne Budd, MEd, MPH

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2007

APPROVED:

Susan F. Franks, Major Professor
Richard Herrington, Committee Member
James Hall, Committee Member
Jose Toledo, Committee Member
Linda Marshall, Chair of the Department of Psychology
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies

Budd, Margaret Anne. *Boston Naming Test with Latencies (BNT-L)*. Doctor of Philosophy (Health Psychology and Behavioral Medicine), May 2007, 168 pp., 15 tables, 9 figures, references, 129 titles.

Although most people have experienced word-finding difficulty at one time or another, there are no clinical instruments able to reliably distinguish normal age-related effects from pathology in word-finding impairment. Two experiments were conducted to establish a modified version of the Boston Naming Test (BNT) that includes latency times, the Boston Naming Test of Latencies (BNT-L), in order to improve the instrument's sensitivity to mild to moderate word-finding impairment. Experiment 1: Latency times on the 60-item BNT (Goodglass et al., 2001) for 235 healthy adults' ages 18-89 years were collected on a representative sample. Qualitative features of the BNT items, statistical analyses, IRT, and demographic considerations of age, gender, education, vocabulary, race and culture, helped create a reduced BNT-L version with 15 of the most discriminating items. Statistically sound and sophisticated normative tables are provided that adjust for unseen covariates. Response latencies did not indicate earlier age-related decline in an optimally healthy sample. Experiment 2: Twenty-three patients referred for neuropsychological testing were administered the BNT-L. Patients referred for evaluation of mild cognitive impairment or possible dementia produced significantly different response BNT-L latencies from the healthy sample whereas patients referred for mild brain injury evaluation did not. Normal word-finding problems were discussed in terms of serial stage models of lexical access, as well as in terms of automatic and controlled cognitive processes in younger and older adults. Statistical process for creating a psychometric instrument using latencies is illustrated.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The ability to represent objects with names provides the basis for human language. Referring to things by name is, by and large, an automatic process people typically take for granted unless something falters and we cannot access the right word at the right time. When this happens we are often certain that the word is within our memory, that the word *is* present, but we are unable to access it, maybe temporarily or maybe indefinitely. Questionnaire studies (Reason, 1984) indicate that word-finding problems occur regularly with most people and that healthy older people report more frequent difficulties with word finding in everyday activities as age increases (Lezak, 2004; Lovelace & Twohig, 1990; Schmitter-Edgecombe, Vesneski, & Jones, 2000). Some researchers view word-finding problems as a natural part of cognitive aging that does not become clinically or statistically significant until late in life (Albert, Heller, & Milberg, 1988; Nicholas, Obler, Albert, & Goodglass, 1985; Nicholas, Barth, Obler, Au, & Martin, 1997). Other research support a view in which there may be a natural decline with age for a few individuals (Van Gorp, Satz, Kiersch, & Henry, 1986) but that in general, naming deficits are not a universal occurrence with aging because many individuals retain excellent word-finding abilities throughout old age (Cruice, Worrall, & Hickson, 2000; MacKay, Connor, & Storandt, 2005). Compared with other cognitive domains, changes in language skills are often small. If changes are present, the changes are rather subtle. In fact, in healthy aging recognition vocabulary often increases through the 50's (Smith, as cited in Nicholas et al., 1985) and lexical comprehension may not change at all (Burke & MacKay, 1997; Nicholas et al., 1985).

Research points to word-finding problems in conversation as the biggest complaint elders have about the effects of aging on cognition (Nicholas et al., 1997). Ninety-five percent of older adults interviewed by Lovelace and Twohig (1990) reported ever experiencing failure to find a

word in a conversation, and 42% reported to experiencing it weekly. Although word-finding problems have also been found for younger adults and young head-injured adults (Sunderland, Watts, Baddeley, & Harris, 1986), subjective complaints about them increase with age (Lovelace & Twohig; Martin & Zimprich, 2003). The effect that word-finding problems can have on aging individuals is considerable. Social isolation, depression and other consequences may occur when individuals have insecurities or embarrassment (Lovelace & Twohig) about their ability to converse with others. Because word-finding problems could be an early indicator of more serious impairment, such as dementia (Calero, Arnedo, Ruiz-Pedrosa, & Carnero, 2002; Georgieff, Dominey, Michel, Marie-Cardine, & Dalery, 1998; Goodglass, Kaplan, & Barresi, 2001; Loring, 1999; Van Gorp et al., 1986), it is a serious and valid concern for both the individual as well as the clinician.

Wording-finding abilities are measured primarily using confrontational naming tasks (Gordon, 1997; Goulet, Ska, & Kahn, 1994; Lezak, 2004; Lopez, Arias, Hunter, Charter, & Scott, 2003; Nicholas et al., 1997) and the Boston Naming Test (BNT) is the most commonly used instrument of this kind (Calero et al., 2002; Schmitter-Edgecombe et al., 2000; Van Gorp et al., 1986). Administration of confrontational naming tasks most commonly involves presenting a person with a card showing a picture and asking for the name of the object shown on the card. The task requires a person to visually identify the object on the basis of an iconic representation, and then mentally retrieve the correct word, hence it is often referred to generically as a "word-finding" task.

More specifically, these are "naming" tasks. Naming involves associating a concept—generally a concrete object that can be pointed to in the environment—with a specific noun. Because a well formed response to a naming task consists of a one-word utterance involving the

singular form of a concrete noun given in an unmarked citation case, naming tasks are especially useful in linguistic investigations of lexical access and retrieval, where complexity arising from morphological, syntactic and discourse level effects must be controlled for. In the remainder of this paper the term "word-finding" will refer to a person's general ability to produce the appropriate word in a given communicative setting. The term "naming" will refer to a particular type of word-finding scenario in which a subject is prompted to name a visually presented object or picture. Finally, while some attempt is made in Chapter 4 to distinguish between the terms "lexical access" and "lexical retrieval," where "retrieval" denotes successful completion, these terms will otherwise be used interchangeably to refer to the automatic neurolinguistic processes involved in mapping a concept to a particular entry in the mental lexicon.

Many variables can affect word-finding ability in naming tasks – individual variables such as age, gender, education, intelligence, and health status; and, environmental variables such as exposure time of the stimulus, priming effects, and properties of the target word. All of which influences the many cognitive processes involved in efficient speech production.

Despite the numerous studies that have found decreased naming abilities with age (Fastenau, Denburg, & Mauer, 1998; Kent & Luszcz, 2002; LaBarge, Edwards, Knesevich, 1986; MacKay, Connor, Albert, & Obler, 2002; Tsang & Lee, 2003), studies have found no relationship with healthy aging (Hickman, Howieson, Dame, Sexton, & Kaye, 2000; Kent & Luszcz, 2002; Nicholas, Brookshire, MacLennan, Schumacker, & Porrazzo, 1989; Tombaugh & Hubley, 1997) and others have found the decline to be only minor (Van Gorp et al., 1986; Welch et al.). Borod, Goodglass, & Kaplan (1980) were one of the early researchers who found a quantitative decline in naming ability with increasing age in healthy adults. Their method of measuring naming ability was with the BNT. Since then other research with the BNT (Albert,

Heller, & Milberg, 1988; Hodgson & Ellis, 1998; Kim & Na, 1999; Nicholas et al., 1985; Welch et al., 1996) confirmed Borod et al.'s (1980) findings, showing both significant differences in naming among age groups, and observing a sharp decline for individuals in their 70s, or 80s (Kent & Luszcz, 2002). Conversely, others believe neurocognitive functions remain relatively stable over time (Hickman et al., 2000) and naming difficulties, in particular, are not a general trend in healthy aging (MacKay et al., 2005) because many of the oldest individuals continue to score near ceiling levels, and many methodological flaws have been identified in the research involving naming ability and age (see Feyereisen, 1997; Goulet et al., 1994 for reviews of flaws with aging research). In fact, some studies have even found improved naming performances with age in both a normal population (Cruice et al., 2002; Farmer, 1990; Schmitter-Edgecombe et al., 2000) and in a clinical population (Thomson & Heaton, 1989).

Despite the fact that naming is often treated as a straightforward operation (Gordon, 1997), there is actually quite a bit of controversy regarding the precise etiology of naming difficulties. Most cognitive research on word-finding has tended to focus on isolating points of failure during the phase of linguistic processing referred to as lexical retrieval, with special attention being paid to anomia, dysnomia, and tip-of-the -tongue (TOT) phenomenon. While cognitive models of lexical access differ in many specifics, they agree with respect to their framing of the problem. First, it is generally accepted that lexical access involves a fairly circumscribed region of the brain, specifically, the left parasylvan areas (Kemeny et al., 2006). Second, all of the models employ two distinct systems corresponding to semantic and phonological levels of representation linked by a third generally referred to as the "mental lexicon." Finally, it is universally acknowledged that lexical access occurs extremely rapidly,

and that latencies in excess of approximately one second represent a failure of lexical access (Brown, 1991).

Studies that propose an age-related decline in naming abilities generally conclude that naming ability remains relatively stable across the adult life span until late in life. Because instruments commonly used to measure word-finding ability such as the BNT are designed to detect lexical access impairments characteristic of clinical dysnomia, there may be paralinguistic or extralinguistic determinants of word-finding ability that tests of naming *accuracy* effectively filter out. Measuring latency to response for adult age groups may show paralinguistic activity and that changes are measurable at an earlier age. Recording response times has been suggested as being a more sensitive method to estimate the extent of word-retrieval (Dunn, Russell, & Sakina, 1989) even though it is rarely used by clinicians or researchers (Goulet et al., 1994; Tsang & Lee, 2003).

The main purpose of this study is to create a new instrument to assess word-finding ability, using latency response times rather than accuracy alone, that is sensitive enough to distinguish between mild impairment and normal aging. After measuring and analyzing latency response times from healthy normal adults using the 60-item BNT (Goodglass et al., 2001), BNT items will be selected to produce a shortened version of a naming test using latency measures – the Boston Naming Test of Latencies (hereinafter referred to as BNT-L). The end result of this study will not only reassess age-related effects on naming ability through latency data, but will also create a new instrument with ecological validity and clinical utility with accompanying normative data that considers all relevant demographic influences.

CHAPTER 2

WORD-FINDING, ANOMIA & TOT

Word Finding and Anomia

The *INS Dictionary of Neuropsychology* (1999, p. 13) defines anomia as "The impaired ability to name objects or retrieve words." Anomia refers to a pathological word-finding difficulty rather than normal word-finding difficulties or vocabulary limitations. Difficulty with word finding is one of the most common speech production disorders for individuals with neurological pathology and for normal individuals with functional impairments (Geschwind, as cited in Georgieff et al., 1998). A great deal of research into word finding difficulties has been driven by the belief that careful analysis of word-finding failures may provide information concerning the process of lexical retrieval and the structure of lexical storage in healthy young adults (Brown & McNeill, 1966; Mitrushina, Boone, & D'Elia, 1999), how these processes and structures are affected in normal aging (Brown & Nix, 1996; Burke et al., 1991), and their role in aphasia (Kohn & Goodglass, 1985).

"Aphasia" is an "acquired disorder of symbolic language processing" (Loring, 1999, p. 15) characterized by a combination of deficiencies in processes involved with language (e.g., comprehension, fluency, repetition), and includes "anomia." Anomia is observed in virtually all types of aphasia (Goodglass et al., 2001), but *not* all subjects who experience word-finding problems are aphasic (Lambon Ralphe, Moriarty, & Sage, 2002). "Anomia" is the clinical term used when the ability to name is pathologically impaired. Anomia denotes difficulty in saying or writing particular words that are appropriate to the situation (Brookshire, 1971), where the speaker cannot produce specifically sought words either during regular conversation or during naming tasks (Loring, 1999). Goodglass et al. (2001) states there is a "qualitative difference between the general restriction of vocabulary, common to most aphasic patients, and the

selective loss of ability to evoke specific words, which is called 'anomia'" (p.7). Individuals with pure anomia require more time to retrieve a selected word but their comprehension and other language processes are in tact (Lambon Ralphe et al., 2002). Anomia often remains a residual impairment even after considerable neurological restoration has occurred following a brain insult (Dunn et al., 1989). However, it is rare for anomia to be an isolated symptom in aphasia (e.g., "classic anomia") (see Lambon Ralphe et al. for case examples) but it is not unusual for anomia to be an isolated symptom of Alzheimer dementia (Georgieff et al., 1998). The terms "dysnomia" and "anomia" are not synonymous although the terms are often used interchangeably. Dysnomia signifies a less severe naming impairment than anomia (Loring, 1999).

Normal people are considered to experience dysnomia or anomia when they experience the inability to find a word accompanied by the characteristic feeling of having a word on "the tip of the tongue" or TOT experience (Georgieff et al., 1998). Brown and McNeill (1966) were first to empirically define the TOT phenomenon when they demonstrated that individuals without impairment could experience anomia after being presented with definitions of rare words in a word-finding test. It appears that anomia exists on a continuum where, on one end, normal individuals experience occasional *intermittent* TOT "states" and, on the other end, those with aphasic disorders and severe clinical anomia experience a *perpetual* TOT "state."

TOT

R. Brown and McNeill (1966) described TOTs as a "TOT state." The insinuation that the experience is "separate from normal waking consciousness" (Brown, 1991) was deliberately in response to William James (1893), who is recognized as the author of the first published description of the TOT experience:

> The state of our consciousness is peculiar. There is a gap therein; but no mere gap. It is a gap that is intensely active. A sort of wraith of the name is in it, beckoning us in a given direction, making us at moments tingle with the sense of our closeness and then letting us sink back without the longed-for term. If wrong names are proposed to us, this singularly definite gap acts immediately so as to negate them. They do not fit into its mould. And the gap of one word does not feel like the gap of another, all empty of content as both might seem necessarily to be when described as gaps. (p. 251).

This description fits with the subjective turmoil people convey while struggling for the intangible word (Brown, 1991). On a list of 28, the TOT experience was listed as the most frequent memory difficulty among older adults (Sunderland et al., 1986), further illustrating the emotional or agitation that is associated with TOT (see R. Brown & McNeill, 1966).

Research methods to study TOTs have involved definitions (R. Brown & McNeill, 1966), self-assessment questionnaires (Sunderland et al.), diary methods (Burke, MacKay, Worthley, & Wade, 1991, Experiment 1), or laboratory techniques (Burke et al., 1991, Experiment 2). A thorough review of TOTs by Brown (1991) listed consistent findings from TOT research: TOTs increase with age; TOTs occur in all ages, including children; TOTs appear to be universal; a person can guess the first letter of the target word 50% of the time, and often the last letter can be guessed (better than chance), but not the letters in between; and, within one minute following the failure, about 50% of TOTs are resolved. See Brown (1991) for a full review of the TOT experience.

CHAPTER 3

FACTORS AFFECTING NAMING ABILITY

When investigating naming, it is important to consider all of the factors that may influence the ease in which a person finds lexical retrieval. Many factors can affect a person's ability to find the correct word at the correct time. Prior research has isolated several variables potentially related to word-finding ability: age (Albert, Heller, & Milberg, 1988; Farmer, 1990; Kent & Luszcz, 2002; LaBarge et al., 1986; Nicholas et al., 1985; Nicholas et al., 1997; Randolph et al., 1999; Van Gorp et al., 1986; Welch, Doineau, Johnson, & King, 1996); education (Calero et al., 2002; Farmer, 1990; Kent & Luszcz; Kim & Na, 1999; Nicholas et al., 1985; Randolph et al.; Thompson & Heaton, 1989; Welch et al., 1996); IQ (Albert et al., 1988; Thomas & Heaton; Thomas et al., 1977; Van Gorp et al.); health status (Albert et al., 1988; Thomas et al., 1977); memory (Albert et al., 1988; Burke & MacKay, 1997; Schmitter-Edgecombe et al., 2000); verbal fluency (Albert et al., 1988; Calero et al.; Dunn et al., 1989; Goodglass et al., 2001; Thomas & Heaton); stress (Brookshire, 1971); properties of the target word (Hodgson & Ellis, 1998; Le Dorze & Durocher, 1992; Mitchell, 1989; Poon & Fozard, 1978; Thomas, Fozard, & Waugh, 1977); caffeine (Lesk & Womble, 2004), gender (Kent & Luszcz; Kim & Na; Randolph et al.; Welch et al., 1996), and priming (Brookshire, 1971; Thomas et al., 1977). Important findings from the studies listed above can be categorized according to whether they focus on variables intrinsic to the individual or on environmental factors affecting naming.

Individual Variables

The following is a general overview of variables concerning the individual that may affect naming performance. The list is not inclusive. Some of the topics below will be addressed in more detail, along with individual variables concerning race, ethnicity and culture, as they

relate specifically to performance on the BNT in Chapter 9: Individual Variables and BNT Performances.

Age

Even when it is not the primary object of investigation, age is frequently included as a variable in word-finding studies because many suspect that age does have a significant effect (Albert et al., 1988; Borod et al., 1980; Fastenau et al., 1998; Kent & Luszcz, 2002; LaBarge et al., 1986; MacKay et al., 2002; Tsang & Lee, 2003), and there is little dispute that subjective reports of word-finding problems increase with age (Lovelace & Twohig, 1990; Nicholas et al., 1985; Sunderland et al., 1986). In general, a subjective complaint about one's overall cognitive functioning increases with age (Martin & Zimprich, 2003). Although most research indicates little relationship between the level of functioning and subjective complaints in both normal (Martin & Zimprich) and brain-damaged individuals (Ponds, van Boxtel, & Jollies, 2000), age-related declines have been documented in different domains of cognitive functioning (see Smith & Rush, 2006) and speed of information processing (Salthouse, 1996). With respect to cognitive tests that show increased variability in the oldest age groups, Randolph et al. (1999) states "it is unclear whether the increased variance simply represents the greater range of scores available as the mean moves away from the ceiling, or whether the increased variability should be interpreted as indicating that old age can be considered a disease state of sorts" (p**.** 494**).** Before concluding word-finding problems are an inevitable consequence of natural aging, researchers (MacKay et al., 2005) recommend using caution, especially because the results in the literature on naming ability and aging are mixed (Goulet et al., 1994) and appear to be dependent on the research design used (Cruice et al., 2000). The age-related decline found in naming ability may be simply due to slower response times (Thomas et al., 1977), or cohort effects (Cruice et al., 2002), and

not naming impairment. Nonetheless, normative data should determine if an elderly person with a low score on a naming test is showing signs of cognitive impairment or normal aging. A list of recommended cautions and considerations when evaluating BNT research and normative data is presented in Chapter 8: Boston Naming Test. See also specific information relating age to BNT performances in Chapter 9: Individual Variables and BNT Performances.

Gender

The nature-nurture question continues to be an uncertainty when discussing differences between males and females in cognitive abilities (Lezak, 2004). While gender differences have been found in brain anatomy, the effects of socialization and education clearly also play a role (Geary, 1989). The effects of gender on naming performances are mixed, and gender is typically considered a weak variable in relation to naming abilities (Lezak, 2004). However, gender differences observed in performances for specific BNT items are noteworthy. See Chapter 9: Individual Variables and BNT Performances for a discussion of gender as related to BNT items.

Education, IQ and Verbal Ability

Education and IQ

Although there are studies that did not find education to be related to naming performances (Albert et al., 1988; Farmer, 1991; Fastenau et al., 1998; LaBarge et al., 1986; Nicholas et al., 1985), an individual's ability to name common objects may be influenced by education or intelligence (IQ). It is reasonable to expect individuals with higher education to have larger vocabularies and to perform better on naming tasks than less educated individuals. Indeed, several other studies have found significant relationships between level of education and picture-naming abilities (Borod et al., 1980; Hawkins et al., 1993; Henderson et al., 1998; Tombaugh & Hubley, 1998; Kent & Luszcz, 2002; Kim & Na, 1999; Lansing, Ivnick, Cullum, &

Randolph, 1999; Nicholas et al., 1989; Spreen & Strauss, 1998; Thomas et al., 1977; Thompson & Heaton, 1989; Welch et al., 1996; Whitfield et al., 2000; Worrall, Yiu, Hickson, & Barnett, 1995). Welch et al. (1986) suggested that a person's naming ability was retained into the 80s if they had more than 12 years of education.

Cognitive functioning related to intelligence has been associated with naming ability, especially in the case of vocabulary aptitudes (Albert et al., 1988). The higher scores correlated with higher education was hypothesized to exist because better educated people have a larger vocabulary base which increases naming ability (Henderson et al., 1998).

Vocabulary

Hawkins et al.'s (1993) findings best exemplify the influence of lower educational level, and of limited vocabulary in particular, on published norms for the BNT (using version from Kaplan et al., 1983). Hawkins and colleagues discovered high false-positive rates on BNT performances when participants had low reading vocabulary scores. Normal participants in their sample scored nearly two standard deviations below the means published with the BNT norms. The average education level for these subjects was 12 and 13 years, probably representative of the population at that time. This study also reports a strong correlation between a vocabulary test and BNT scores (r = .81, and .83 when illiterate subjects were excluded), which could suggest, "that in some circumstances the BNT essentially measures vocabulary" (Hawkins & Bender, 2002, p. 1143). When applying the norms used in Hawkins et al., caution is especially warranted with individuals with lower-than-average reading ability (Spreen & Strauss, 1998). Other studies have found verbal intelligence, as measured by WAIS-R vocabulary scores, to highly correlate with naming performance on the BNT (Albert et al., 1988; Thompson & Heaton, 1989).

Verbal Fluency

The relationship between age-related changes in verbal fluency and word-finding ability is not clear (Garcia & Orange, 1996), however, a few studies have shown a weak relationship between measures of verbal fluency and word-finding abilities (Albert et al., 1988; Schmitter-Edgecombe et al., 2000). Verbal fluency was used as an apriori variable for naming ability in Albert et al.'s (1988) study, but verbal fluency was not incorporated in the final statistical model of contributors to naming ability after the researchers concluded it was to not a related factor in naming for healthy adults. Schmitter-Edgecombe et al. (2000) did not find a significant correlation between poor performance on a verbal fluency test and poor performance on a discourse test used to measure word-finding abilities. However, this same study found a significant correlation (r = .33) between individuals' performances on a verbal fluency test (category, animals) and BNT naming scores. Similarly, Calero et al. (2002) also found a significant correlation (r = .42) between a verbal fluency test (both semantic and phonemic cues, e.g. conceptually related nouns and words that begin with letter *p*) and BNT scores. As an individual variable, one's verbal fluency may not influence performance on a naming task that requires selection of one specific word (Brookshire, 1997). Verbal fluency tests have been used to assess word finding and more will be discussed in Chapter 7: Measuring Word Finding.

Age of Acquisition

The age in which one learns a word is called the "age of acquisition." The age in which an object's name was first learned affects the vulnerability of that name to retrieval failure (Hodgson & Ellis, 1998). The earlier the age at which a word was acquired, the more robust it is to word-finding failure (Lezak, 2004), and words acquired earlier are more resistant to some types of brain injury than words acquired later (Ellis, Lum, & Lambon Ralph, 1996). Likewise,

later age of acquisition is associated with more errors. Age of acquisition is an individual variable that directly affects one's speed and ability to name a target word (Barry, Morrison, and Ellis, 1997; Hodgson & Ellis, 1998). This is discussed in more detail in a separate section, Chapter 5: Response Latencies, section titled Properties of Object Names.

Memory

When older people complain about memory problems, they are often referring to reduced word-finding ability (Burke & MacKay, 1997; Lezak, 2004). While some researchers have speculated as to whether word retrieval in the case of naming and other memory subtypes are actually very different (Albert et al., 1988), others presented a clearer distinction between memory functions and naming functions (Lezak, 2004). Many systems are apparently involved with both naming ability (Lambon Ralph, Moriorty, & Sage, 2002) and memory functions (Lezak, 2004) and deficits in processes outside these systems can affect either naming ability or memory functioning, or both. For example, attention and concentration are processes outside the system for naming and the system for memory, yet deficits in attention and concentration can affect performances in one or both domains. An example more specific to naming, a person can have difficulty recalling episodic memories but not have difficulty in retrieving common words or names consistently (Lezak, 2004). When conceptualizing test findings and theory, maintaining terminological distinctions between aspects of a particular function (e.g. naming) and other functions necessary for efficient functioning (e.g. episodic memory) can help dissociate two related functions.

Specifically to BNT research, Schmitter-Edgecombe et al. (2000) found little evidence to support word-finding problems was worsened by poorer memory. Schmitter-Edgecombe et al. study found no correlation between two memory measures (list learning and Delayed list

memory) and two word-finding measures; BNT naming scores and discourse test performance was not related to these measures of memory. On the other hand, in the same study verbal fluency scores were related to these measures of memory.

A general processing theory called "new connection formation" helps clarify some of the findings in the literature relating memory, aging and language. This theory states that declines in memory systems with aging occur only when "new connections," or new formations, between memory representations are required, and that existing memory systems are spared (see Burke & MacKay, 1997). For example, episodic memory system (ability to remember events situated over time and place; e.g. placed keys) has been deemed a separate memory system and one at risk in aging (Mitchell, 1989). However, closer inspection of studies shows an age-linked decline primarily happens to new or recent events (or laboratory experiences), and no difference between aging occurs when retelling past events or experiences at a younger age. The new connections theory of memory and aging challenges the multiple memory systems theory that is often included in discussions on age effects and memory. For example, multiple systems theory could suggest episodic memory declines with age, but semantic memory is stable. The new connections theory would state that both types of memory are stable for existing memory representations, but age changes occur with the formation of *new* memory representations. New semantic information (Burke & MacKay, 1997), for example, shows typical age-related declines. The new connections theory could explain Hickman et al.'s (2000) overall finding of neurocognitive stability in age despite an observation that older participants in their longitudinal study did not exhibit practice effects like the younger participants. The new connections theory has been empirically demonstrated with episodic, explicit memory, and semantic memory (Burke & MacKay, 1997).

Naming is most often thought of as involving semantic memory. *The INS Dictionary of Neuropsychology* (1999) defines semantic memory:

> Memory that is context-free, reflecting general knowledge of symbols, concepts, and the rules for manipulating them. In contrast to episodic memory, semantic memories rarely concern specific information about situations in which they were learned … Impairments in in semantic memory generally do not occur unless there is an acute confusional state, dementia of at least moderate severity, or focal lesions affecting specific aspects of linguistic function. (p. 105)

As implicitly stated in the definition, semantic memory is rather robust and remains relatively unimpaired with natural aging. Tests of general knowledge (Nyberg, Backman, Erngrund, Olofsson, & Nilsson, 1996) or vocabulary (McGurn et al., 2004) show age consistency through adulthood. This supports our hypothesis that no age-related decline will be observed with the semantic aspects of word retrieval, but that age changes may be reflected in other aspects of word-retrieval that may be indicated by measuring latencies on naming tests rather than by measuring accuracy alone.

Health

Issues of poor health complicate attempts to understand aging effects on cognition, especially when older individuals are more likely to have chronic medical problems than younger individuals (Hickman et al., 2000). Poor control for the health of the participants has been considered a possible source for the mixed results in the literature concerning aging and naming (Albert et al., 1988; Goulet et al. 1994; Kent & Luszcz, 2002). Some studies recognized this potential to confound, especially in an older cohort, and attempted to control for it by using only "optimally healthy individuals" in their attempts to investigate differences in naming

ability. In each of these, the age differences in naming performance remained (Au et al.; Hickman et al.; Randolph, 1999; Whitfield et al., 2000). MacKay et al. (2005) ruled out dementia as an explanation for decreased BNT scores in older adults. Whitfield et al. (2000) examined health status, health habits, physical functioning, and speed of performance and BNT performance and found three predictors of BNT scores for European Americans: fewer reported symptoms of depression, higher peak expiratory flow, and smoking. An individual's health status may affect overall cognitive abilities, especially with elderly populations. The heterogeneity of health levels in an elderly population must be considered when assessing this population and using normative data. This is necessary in order to discern whether changes in cognitive functioning are due to disease processes or to the aging process itself.

Stress

Using a naming task other than the BNT, Brookshire (1971) concluded that stress was a factor for differences in naming ability in his participants. Specifically, he found that the exposure time of the stimulus affects performance in anomic individuals. Shorter exposure intervals of the stimulus were inferred to create stress which interfered with naming performance. Participants performed best when they were able to pace the exposure time of the stimulus.

Environmental Variables that may Affect Naming

Exposure Time of Stimulus

When investigating stimulus exposure time and age effects, Thomas et al. (1977) found that older participants required longer presentation times to correctly identify the picture stimulus than younger participants. For example, Thomas et al. found on average, 19-26 year olds needed 84 msec (0.084 seconds) and 56-74 year olds needed 115 msec (0.115 seconds) to correctly

name a picture. Interestingly, several studies have found significant correlations between naming accuracy and stimulus presentation time even when measured using timescales using increments several orders of magnitude longer than that used by Thomas et al. In addition to finding anomic patients had the best naming performance when they were able to self-pace the time of stimulus exposure, Brookshire (1971) recorded exposure times in normals to determine the least amount of exposure time needed for a correct naming response. Most subjects named an item correctly with presentation of the stimulus for 10 seconds, and performance slightly improved for some subjects when the stimulus was presented for 30 seconds. Similarly, using a technique in which the subject controls the presentation time (up to 15 seconds) by page turning, Hodgson and Ellis (1998) reported younger adults provided more correct responses to a naming task with stimulus presentation in the range of 0-5 seconds than older adults, and older adults responded correctly with presentations in the 5-10 second range more than younger adults. Overall, these results show that elderly people had less accuracy and required longer presentation of the stimulus to name objects compared to younger individuals. None of the studies above used the BNT to obtain their findings.

Most studies using the BNT have not placed additional limits on the duration of stimulus presentation, either presenting the stimulus for an unlimited time (Albert et al., 1988) or allowing 20 seconds per test item (Spreen & Strauss, 1998). Based on latency studies not using the BNT, it appears that the effects of different exposure time may not be significant if stimulus exposure was at least 15 seconds; exposure time to stimulus cards in previous research has shown no affect on response times or accuracy after 15 seconds of exposure to the picture (Brookshire, 1971).

Priming

The semantic priming paradigm is the most common technique for assessing the processes of word organization and meanings in semantic memory (Burke & MacKay, 1997). The semantic priming paradigm refers to the reduction in time needed to state a target word due to a semantically-related or semantically unrelated word preceding it. For example, the target word *doctor* may be identified quicker following a semantic prime *nurse* as opposed to an unrelated semantic prime *chair.* Priming is considered an automatic process that is not under conscious control. Support for the existence of priming effects on word finding is generally available for normals (Burke et al., 1991) and anomics (Lambon Ralph et al. 2000), but one study did not find naming practice helpful in improving naming performance in individuals with aphasia (Brookshire, 1971). Lambon Ralph et al. (2000) demonstrated both the positive and negative effects of priming on individuals with classic anomia. In this study, Lambon Ralphe and colleagues demonstrated that strategic priming is effective and could either make anomia better or worse. Anomia was made better by facilitating resolution of a TOT with repetition priming or by providing the first phonetic cue. Anomia was made worse by suppressing naming by providing an incorrect phonemic cue.

No findings support priming effects change with age, suggesting a basic integrity of language comprehension in aging (Laver & Burke, 1993). Priming effects also decreased the effects of aging in one study (Thomas et al., 1977). These two findings would support the hypothesis that automatic lexical processing is not affected by aging, and that priming would help facilitate controlled processing which could level off any age differences.

Properties of Target Word

The properties of the target word (i.e. word length or word frequency) have known effects on naming latency for picture tasks (Hodgson & Ellis, 1998). For an overview see section "Properties of the Objects or their Names" located in Chapter 5: Response Latencies.

Summary

Naming accuracy and latency can be influenced by factors that are considered individual attribute variables, such as age, gender, education, IQ, verbal abilities, memory abilities, and the age at which target words were acquired. Other factors may be dependent on a particular state, such as poor health, or are extraneous, such as stress. Environmental or purely external variables surrounding exposure time of the stimulus, semantic and phonological priming, and word properties could also impact naming abilities. As discussed, many variables should be considered when evaluating the effects of age on word-finding ability (Schmitter-Edgecombe et al., 2000). Failure to name an object may be due to many cognitive factors or many environmental or external variables. The research on some of the factors that could affect naming ability is mixed and on other factors the findings are indisputable. There are questions whether age-related effects, such as memory or verbal ability, could impact word-finding ability as well as the basic integrity of the language system. Knowledge about variables that influence naming abilities is paramount in the interpretation of test results, and whether or not the factors can be controlled.

CHAPTER 4

COGNITIVE MODELS OF NAMING

Background

In ordinary conversation, it is generally estimated that people produce words at the rate of about two to four a second (Levelt, 2001). This feat is performed unconsciously and without effort except in the extremely rare case where one is suddenly unable to "find" the right word. [1] The ease with which human beings are able to use words tends to mask the real complexity of the task of referring to things--a task that has intrigued linguists for literally thousands of years.

In the twentieth century the problem of reference has been vigorously investigated in fields as diverse as philosophy, computer science, neuroanatomy and the various branches of linguistics. One result of this proliferation of research from such a variety theoretical perspectives has been to confuse a great deal of the common terminology. Therefore, before discussing cognitive models of word finding, it is necessary to first clarify what words are generally thought to be.

What's in a Name?

First, words are symbols, that is, they refer to something other than themselves. A natural impulse might be to say that words refer to objects; "cup" refers to the object from which one drinks liquids like coffee. Obviously, though, the world contains more than one physical object suited to this purpose. Furthermore, words like "unicorn" refer to things that do not strictly exist, so it is more accurate to view words as referring to mental categories or "concepts" rather than things in themselves. The linguistic term for a word's conceptual referent is its semantics.

---

[1] The field of linguistics is often said to have originated with the Sanskrit Grammarians, c. 400-700 BCE.

Second, as Ferdinand de Saussure is famous for having pointed out, a word's form is arbitrary with respect to the concept to which it refers (Saussure, 1986). If, on a whim, a three year old decides that "snigleygoo" means "tomato," once the members of her family have learned the new word, they are free to use snigleygoo at will in place of tomato. In other words, the form of a word can not be inferred from its meaning and vice versa. The question this begs is what constitutes a word's form?

It is common to think of words as having a particular spelling, or to point to groups of letters delimited by white space on a printed page as examples of words. This is misleading, though, since writing is an invented technology used to store words in a non-volatile form, not an innate ability like walking and talking.[2] Likewise, overt marking of word boundaries with white space has no acoustic analog (pauses between words) in fluent speech. The fact that a spoken utterance consists of a single unbroken stream of sound in which discrete words cannot be isolated solely on the basis of their acoustic properties also means that it would be incorrect to characterize word forms in terms of their manifest acoustical contour (as measured using sound spectrograph). Instead, linguists employ the notion of phonemic representation, or phonology.

The formal definition of a phoneme is that it is a mental representation of a contrasting segment in a given language (Spencer, 1996). In other words, a phoneme refers to mental representations of speech sounds and not to sounds themselves. A phonological segment is the

---

[2] Those not acquainted with the historical controversy surrounding this assertion should see Noam Chomsky's *Review of B. F. Skinner's Verbal Behavior* (1957), or for a less vitriolic presentation of the case for innateness, Steven Pinker's *The Language Instinct* (1994). Linguists generally consider the issue settled.

basic building block from which language is constituted; for segments to contrast means that they are distinguishable from one another on the basis of their phonetic (acoustic/articulatory) expression in a given phonetic context. For example, consider the phoneme /s/ used as the plural suffix "-s" expressed phonetically as either [s], [z] or [əz] depending on its phonetic context as illustrated in the following examples:

"cat" + "-s" [kæts]

"cow" + "-s" [ka□z]

"fox" + "-s" [f□ksəz]

While the phoneme /s/ is expressed variously as either the sound [s], [z] or [əz] depending on the sound that precedes it, the segments /s/ and /z/ constitute separate phonemes since they occur in contrasting distributions; they are distinguishable when they occur in identical contexts: "sip" expressed [s□p] and "zip" expressed [z□p]. It may seem like nit-picking to distinguish so elaborately between the sound of a word and what a word sounds like to a person, but it is important to acknowledge here that the term phome refers to a mental representation at the perceptual level. The phonology of a language, therefore, refers to the inventory of phonemes available to that language, each of which represents an articulatory program that produces a distinctive pattern of phonetic expression in overt speech.

In practice, the term phonology is frequently used more generically to include supra-segmental phonology--syllabification and stress--otherwise referred to as prosody, and even, as is often the case in neurolinguistics, as a sort of shorthand for the whole gambit of linguistic processes involved in mapping an item in the "mental lexicon" to a corresponding articulatory gesture.

Words, therefore, can be viewed as entries in a "mental lexicon," each having a form, a mental representation at the phonological level, and a meaning, or mental representation at the semantic level. The following discussion of models of lexical access and naming presents a distillation of the common aspects of a variety of such models. In order to avoid unnecessary confusion that might result from small inconsistencies in terminology surrounding the mental lexicon itself, the lexical level will occasionally be referred to simply as the post-semantic pre-phonological level of representation.

Lexical Access

The basic picture of the word finding process that emerges from this general notion of what words are is one involving three levels of representation: a semantic representation specifying a concept, a phonological representation specifying an articulatory program, and an intermediate lexical representation that maps a particular set of semantic features to a set of grammatical features (including syntactic and morphological properties) referred to in most neurolinguistic research as its phonological representation. Clarke, Johnson and Pavio (1996) note the close correspondence between this idea of word finding and our intuitive sense of our own volitional capacities with respect to naming familiar objects. They remark that while *recognizing* an object is essentially involuntary, there is some choice involved in deciding what to call it, and likewise, it is certainly possible to know what something is called without actually uttering its name.

Cognitive models of word finding attempt to specify a biologically plausible mechanism by which the brain maps a lexical concept (a concept or bundle of semantic features for which a lexical item exists) to a corresponding phonological form. Many of these models are based on studies of picture naming, a task for which many complicating factors, such as syntactic and

morphological complexity found in fluent discourse, can be controlled. One representative model is Levelt's popular Two Stage Theory of Lexical Access (Levelt, 2001).

Levelt's model, based largely on stimulus onset asynchrony (SOA) studies of speakers' word production latencies, presents a naturalistic neural network able to map a given lexical concept to a corresponding articulatory gesture. Much of the evidentiary support for the model comes from the power of A. Roelofs' WEAVER++[3] implementation of the model to predict relative changes in response latencies corresponding to the co-presentation of various distracter stimuli.

Levelt's network comprises two distinct subsystems, "lexical selection" and "form encoding," which operate in series. The lexical selection network consists of two strata. Nodes in the uppermost strata represent "lexical concepts," concepts for which the mental lexicon contains a corresponding "lemma" or syntactic description. The second strata consist of nodes representing the lemmas themselves.

In the preparatory phase known as "perspective taking," a subject begins to "focus on a concept whose expression will serve a particular communicative goal" (Levelt, 1996, p. 13465). This results in coactivation of semantically related lexical concept nodes which in turn spread activation to corresponding lemmas. The time required for lemma selection to occur depends on the amount of coactivation from related concept nodes, and the target lemma is said to be "selected under competition" (Levelt, 2001, p. 13464). Coactivation of conceptually related lemmas accounts for semantic priming effects. In the second stage, the “form encoding,” only the selected lemma begins to spread activation to the phonological nodes of the form encoding

[3] WEAVER++ is a programming language used to describe neural networks.

network specified for that lemma. With the activation of the appropriate phonological nodes, the form encoding stage proceeds with "incremental syllabification," "phonetic encoding" and "articulation."

In addition to serial two-stage models such as Levelt's, there are also cognitive theories of lexical retrieval that employ interactive-activation models, as well as cascade models involving parallel distributed processing (PDP) principles (Ralph, Sage, Roberts, 1999). The salient feature of all of the models with respect to clinical naming tests such as the BNT, as well as to this present study, is that they all support categorizing failures in lexical retrieval as resulting from either semantic deficit, post-semantic pre-phonological deficit, or phonological deficit. This provides the theoretical rationale for offering semantic cues during administration of the BNT in order to reduce false positives resulting from conceptual mischaracterization of the objects depicted in the test items, as well as for supplying phonological cues following missed items in order to clarify the nature of the linguistic deficit.

Other Tasks Involved with Naming

Despite the fact that investigations into the nature of word finding typically employ naming tasks to provide a window on the process of lexical retrieval, it is important to remember that lexical retrieval *per se* is a relatively brief component in the process of naming a pictured object. Before retrieval of a lexical item can proceed, the subject must visually identify the object and, in the case of picture naming, decipher what object the image depicts. A subject who fails to recognize the object depicted will not only be unable to name the object, but will be unable to explain what it is used for, or to produce semantically related words. Once recognition and interpretation of an image succeeds and lexical retrieval takes place, subjects may hesitate before

articulating a response. Figure 1 depicts a more complete serial stage model of naming along with the anatomical regions involved at each stage.

There is general agreement that to successfully name a picture, the following cognitive operations must take place: visuoperceptual processes, object recognition and semantic processes, lexical processes, and articulatory processes (Barry et al., 1997; Nicholas, et al., 1997). Visuoperceptual processes involve the ability to see and recognize the item. Perceptual aspects generally are not considered to be large contributors to word-retrieval difficulties (Hodgson & Ellis, 1998), however, some researchers believe perceptual problems in older individuals could account for some of the word-finding difficulties in picture naming tasks (Thomas et al., 1977). As noted in the discussion of stimulus presentation time as a factor in the performance of naming tasks presented in Chapter 3 above, the finding (in Hodgson & Ellis, 1998) that increased stimulus presentation times (up to 15 seconds) correlates with improvement in older subjects' naming accuracy (Hodgson & Ellis, 1998) can be seen to support the view that visual perception and object recognition may be more of a factor in naming than is often assumed. Additionally, object recognition is highly influenced by "image agreement," or how the picture (image) matches (is in agreement) with the rater's mental image of the object. (A related topic, Object Familiarity will be discussed in the next chapter). Barry et al. (1997) were the first to use image agreement as a variable and found pictures rated highly on image agreement were named more quickly than pictures with less image agreement.

Processing Models

Stern, Prather, Swinney and Zurif (1991) apply two discrete processing models in their treatment of naming: automatic processing and controlled processing. "Automatic processing," assumed not under control of the subject, is not affected by intention or attentional processing.

Automatic access is fast-acting (300-700 msec), and is typically what occurs when lexical retrieval is successful. In contrast, "controlled processing" places demands on processes of attention and is affected by intention, or by use of cognitive strategies. Controlled effects occur when lexical retrieval fails and a person actively searches under conscious control. This begins "post-lexical entry" and can likely be associated with a TOT experience, or when a person is consciously using strategies to locate the correct word. Automatic processing is diminished after 1100 ms (Stern et al.) and subsequent effects can be attributed to controlled processes. This coincides with Dunn et al.'s (1989) finding that the average response time to name a picture was 1.145 s. Responses after this amount of time suggests controlled processing when typical lexical retrieval mechanisms fail and the person is forced to utilize other cognitive processes to access the word.

No age effects were found with automatic processing, routine language processes did not seem to slow with aging in a study conducted by Stern et al. (1991). However, the same study noted age effects were noted in controlled processing, presumably due to the limited processing resources on attention and other cognitive demands involved with aging. Therefore, the locus for age-related slowing found in TOT studies (Brown & Nix, 1996; Burke et al., 1991) and in latency performances on naming tasks (Hodgson & Ellis, 1998; Tsang & Lee, 2003) suggests that word-finding difficulties do not arise in the early, language-specific processing that mediates lexical access, but in later language processing that requires the use of problem-solving strategies and other cognitive resources. This suggests the possibility that the age-related declines in naming latencies and accuracies reflect not age-related declines in lexical retrieval, *per se*, but to differences in strategies and controlled cognitive processes subsequent to the lexical lookup failure.

Conclusion

As mentioned at the beginning of this section, ordinary word finding proceeds at an extremely rapid rate. The latencies involved in SOA studies upon which models of lexical access are often based, for example, are measured in milliseconds and are significant at timescales involving tenths, or event hundredths of a second. The time period during which lexical retrieval either succeeds or fails, therefore, is completely swamped by measurements in whole seconds such as the present study proposes to apply to administration of the BNT, and too brief to be measurable in a typical clinical setting. It is the fact that more general, extra-linguistic faculties must account for the great majority of the twenty to forty seconds allotted for each BNT / BNT-L item that has led to the hypothesis that response latencies for items on the BNT-L may provide a better indication of general cognitive deficits than the accuracy scores alone, which measure only the frequency with which deficits occur in a relatively constrained region of the brain.

CHAPTER 5

RESPONSE LATENCIES

Only a few studies have included latency times when using a picture naming task to assess word-finding abilities (Brookshire, 1971; Dunn et al., 1989; Hodgson & Ellis, 1998; Thomas et al., 1977; Tsang & Lee, 2003). Increased latencies have been related to age for several word production tasks – reading aloud written words, answering questions, and picture-naming tasks (see Amrhein, 1995 for a review of speeded picture-word processing research). Unfortunately, most research on naming ability focuses only on accuracy scores. Goulet et al. (1994) states that accuracy scores are used over latency scores in most naming studies simply because accuracy scores are what has been most frequently used, they are easily available, and are clinically useful. Furthermore, no norms are available for latency responses on picture-naming tests (Goulet et al.; Tsang & Lee). Availability of a clinically useful instrument with normative data on latency times on naming tasks might well spawn more research on this topic, especially since speed of word finding is the complaint stated by most often by the elderly (Lovelace & Twohig, 1990). TOTs would not be a bother or embarrassment if subjects were able to resolve them quickly.

Properties of Object Names

Elderly people often struggle to name some objects in a naming task, while easily naming others. Variables within the individual or in the environment have already been described. The properties of the names given to presented objects have been investigated to help identify some of the causes of naming problems in the elderly. One property relates to an individual variable and the others are external variables; all of these can help shed light on age differences found among naming performances. Five common properties have been investigated in this respect: 1) age of acquisition, 2) word length, 3) name agreement, 4) word frequency, and, 5) object

familiarity. The effects of each property will be discussed individually; however, considering them in isolation may be misleading because many of the properties are intercorrelated (Hodson & Ellis, 1998) and disentangling which property is the operative factor is difficult. For heuristic purposes, a brief explanation of each property will be presented in relation to age and its effect on speed of naming, and where possible, with respect to levels of automatic and controlled processing.

Age of Acquisition

As stated previously, the age at which an object's name was learned affects the vulnerability of that name to word-finding malfunction (Hodgson & Ellis, 1998; Lezak, 2004). The earlier the age at which a word was acquired, the less likely it will produce word-finding failure (Lezak, 2004), and words acquired later in life (which are often longer, less common words) are associated with more failures.

Age of acquisition emerged from regression analysis as the most robust of three independent predictors of naming success of many variables investigated by Hodgson and Ellis (1998) in a picture-naming task. Age of acquisition produced the highest raw correlation with naming accuracy and displayed the highest ability to predict correct naming for all correct responses made within the first five seconds as well as within a 15 second response range. The other two independent predictors of naming success in this study were "word length" and "name agreement" which will be discussed in the next sections.

Age of acquisition can result in a cohort confound by systematically affecting naming scores of younger people differently from older people. For example, Schmitter-Edgecombe's (2000) study found age-related effects on naming ability, however, the validity of these effects is questionable due to a cohort effect that had nothing to do with naming ability. In this study the

majority of younger participants systematically missed four items on the BNT because they did not recognize the target words (yoke, trellis, palette, and abacus), whereas the majority of older participants named the same four items correctly. The authors concluded that these four items had an age bias in favor of older individuals. Part of this bias, or cohort effect, is likely due to when these words were learned (i.e., acquired) for older versus younger adults. Part of this bias in cohort is likely due to when these words were acquired for older versus younger adults.

Age of acquisition has been an important determinant of picture-naming latency (Barry et al., 1997). Morrison et al. (1992) discovered that age of acquisition does not affect object recognition or object identification, but affects object naming. With the sequential processes required to name a picture, this finding suggests that the locus of effect for age of acquisition is at the post-semantic level of processing. Normative data is currently available for age of acquisition for pictured objects; Morrison, Chappell, and Ellis (1997) compared measures of determining age of acquisition and using 220 children reported a set of age of acquisition norms for 297 pictured objects (232 of which came from Snodgrass & Vanderwart, 1980).

Word Length

The effect of word length, or the length of the target word, on elderly people's ability to name objects remains ambiguous by word length's effects on other variables. Intuitively word length is highly correlated with other word properties that could affect naming performances. For example, one would expect for shorter words to be more common than longer words (word frequency) and more familiar (object familiarity) and for shorter words to be learned at an earlier age (age of acquisition). Hodgson & Ellis (1998) confirmed this intuition with significant correlations between length of word and all aforementioned word properties in an elderly population. As mentioned above, word length emerged from Hodgson and Ellis' regression

analysis as a significant independent predictor of picture naming. Longer word items in this same study were named less accurately in both younger and older participants in the 0-5s range. Word length as an independent predictor, however, remained significant only when responses were produced in the 0-5s response range. After 5s, response accuracy was unaffected by word length.

Using models of automatic and controlled processing, the findings above suggest that word length is influential in the early, language-specific process that mediates word retrieval. Continuing in the same model, word length would likely have a lesser effect on processes following lexical failure, when a person is consciously trying to access the correct word. Brookshire (1997) would, however, rebut the application of word length's affect on automatic processes, but not on controlled processes. Brookshire views word length as contributing to the overall complexity of articulation, and more on the mechanical properties of naming functions, whereas he views other properties (e.g., word frequency or word familiarity) as being more likely to affect word access and retrieval functions.

Regarding age and word length, Le Dorze and Durocher's (1992) found an interaction between age and naming while investigating the effect of the number syllables in a target word in young, middle-aged and elderly participants' naming accuracy. Older participants had more difficulty with longer names than younger participants. It should be noted that "word length" can be measured in different ways. The fact that some measure word length by number of syllables (Brookshire, 1997; Le Dorze & Durocher, 1992) and others by the number of phonemic segments in a word (Barry et al., 1997) makes this type of research difficult to interpret.

Name Agreement

Name agreement refers to the "codability" of an object, whether an object can be referred to with another name. For example, "chair" is an object said to have high name agreement

because it has few, if any, plausible alternative names. An object with low name agreement is one which possesses several possible names. For example, a "sofa" can also be called a "couch" or "settee." Studies have shown that objects that have low name agreement, or have more than one potential name, result in slower naming times (Vitkovitch & Tyrrell, 1995). With regards to aging, Mitchell (1989) found no interaction effects between age and name agreement on naming latency; overall, naming latencies were slower in both younger (ages 19-32) and older groups (ages 63-80) when there was low name agreement, with the effect being equally strong in both age groups.

Name agreement was the third independent predictor of naming accuracy found in Hodgson and Ellis' (1998) regression analysis on word properties' effect on naming speed and accuracy in the two time ranges--name agreement was predictive for accurate responses in both the 0-5s range and 6-15s latency range. Intuitively, name agreement would appear to influence both automatic word-retrieval and conscious problem-solving strategies during attempts to find the right word.

Word Frequency

Word frequency is the number of times a particular word is used in common communication. It has been stated that age differences may rely on word length and word frequency (Feyereisen, 1997). Word frequency is similar to word length in that it is highly correlated with other properties discussed. Hodgson and Ellis (1998) found word frequency to be significantly correlated, in descending order of strength of correlation, with: age of acquisition, naming accuracy at 5s latency, naming accuracy at 15s latency, imageability, visual complexity, and name agreement. Word frequency was not as powerful a predictor as other word attributes. In fact, word frequency was often not a factor in picture naming speed at all in several studies

once age of acquisition was accounted for (see Barry et al., 1997 for a discussion). However, a significant interaction of frequency and age of acquisition on picture-naming speed exists: high frequency and early acquisition produces the fastest speeds, and low frequency words and late acquisition generates slower speeds (Barry et al., 1997).

One of the few studies measuring naming latency, Thomas, Fozard, and Waugh (1977) considered word frequency in their assessment of the effects of age on speed of retrieval in a picture naming task. Older participants from an age range of 25 to 74 years produced longer latencies in naming a picture, but the effect of word frequency on naming was the same for both younger and older participants; no interaction was found for word frequency and age on latency to name a picture in this study. The effects of age of acquisition on naming ability were not considered because they were not known at the time of this study.

Though words with low frequency were the target stimuli for the original study of TOTs (Brown & McNeill, 1966), low frequency words have not produced consistent results in TOT experiences (Brown, 1991). For example, Yaniv & Meyer (1987) found a high rate of TOTs in their experiment despite using higher frequency words. TOTs are not restricted to rare words (Brown & Nix, 1996).

The effects of word frequency seem most relevant when assessing culturally diverse populations (Cruice et al., 2002) where word frequency probably correlates less strongly with many other variables. In general, accurate measures of word frequency are not always available (Barry et al., 1997) and therefore, results obtained on this word property should be interpreted conservatively.

Object Familiarity

A classic study on naming and aging, Poon and Fozard (1978) examined naming latency of four categories of objects based on their familiarity to young or old participants. One category of four showed no age differences: naming latencies on "common contemporary" objects (objects used throughout the century pictured in their current form, e.g., phone) did not differ between young and old participants. Older participants were significantly faster on naming objects from two categories that reflected generational familiarity: "common dated" objects (objects used throughout the century but in their dated form, e.g., old camera) and "unique dated" objects (objects that were commonplace when the older participants were younger, e.g., bed pan). Along the same line, younger participants were faster to name the category reflecting their generation, modern objects that arrived during the current decade of the study (e.g., calculator).

Based on Poon and Fozard's findings, it could appear that both age of acquisition and object familiarity can account for some of younger participant's poor performance on the four BNT items found in Schmitter-Edgecombe's (2000) study previously mentioned (yoke, trellis, palette, and abacus). Further supporting this hypothesis is the highly significant correlation (r = -.498) between age of acquisition and object familiarity determined by Hodgson and Ellis, 1998.

Having object or name familiarity does not preclude one from having difficulty finding the correct word, however. Words that caused a TOT but were later resolved were rated as being more familiar than less familiar in Burke et al.'s (1991) diary study of TOTs.

Conclusion

It must be stated again that the true influence of specific word properties on naming accuracy is unclear. The extent to which variables genuinely affect naming accuracy is muddled

when the predictors themselves are intercorrelated, as shown above. Single correlations of some variables may simply reflect their correlations with other variables which have a true influence on naming. Most commonly, however, the speed and accuracy of a confrontational naming test has been associated with both the frequency with which the name of a given object is used (Snodgrass & Vanderwart, 1980) and the age at which the word was acquired (Barry et al., 1997). It makes sense that word frequency would affect naming latency, with commonly used words more accessible as a result of repeated activation. Like many of the properties discussed, word frequency and age of acquisition are known to be correlated (Barry et al., 1997), and are expected to be likewise related to naming performance for unilingual (Randolph et al., 1999), and especially for bilingual individuals (Roberts et al., 2002). In spite of these findings, further analysis of several studies found word frequency had no effect on naming speed once the effects from age of acquisition were removed (See Barry et al., for a review of the studies). Generally, items with a low age of acquisition (Morrison et al., 1992) and high frequency (Cruice et al., 2002) consistently produce the most rapid and accurate responses (Hodgson & Ellis, 1998). Some of the age differences found in naming performances are likely attributable to properties of the target word (Feyereisen, 1997).

CHAPTER 6

LANGUAGE AND AGING

Burke and MacKay (1997) divide the effects of aging on language into two segments: "The Input Side" and "The Output Side" (p.8). The following will illustrate the proposed hypothesis that the affects aging has on word finding is not because of semantic or lexical retrieval failure (semantic level and prephonological level are in tact), but due to post-lexical access (the phonological connections mapping the concept to the point of articulation), processes that utilized controlled processing mechanisms.

Comprehension and Semantic Meaning

The input side refers to the processes of perceiving letters and speech sounds that comprise words and comprehension of the meaning of words and sentences. These skills are robust throughout aging despite sensory deficits (Madden, 1988) and encoding deficits. Studies using the semantic priming paradigm show that older people have the same automatic activation that younger people have following a semantic prime, thus, age differences are not perceptible in the receptive or comprehensive part of speech but differences are noted in the productive aspect (Burke & MacKay, 1997).

Word Production

The "Output Side" includes language production. While language comprehension has shown resistance to aging (Burke & MacKay, 1997), language production has not (see sections on BNT and aging). As previously stated, older adults frequently complain about not finding the right words (Sunderland et al., 1986). The word they seek is a word in which they know, with no deficit in forming an idea to be expressed. Rather, word-finding problems reflect a problem in mapping the well-defined concept onto its phonological or orthographic form. For example, often in a TOT state a person can describe a word and its meaning, generate alternative words

(Cross & Burke, 2004), and can even produce phonological features (Brown, 1991; Lambon Ralphe et al., 2002), yet they cannot generate the desired word.

There were four empirical findings that indicated an age-related decline in word production. One, was that the frequency of TOTs occurred more often as one ages (Brown & Nix, 1996; Burke, et al., 1991). Second, numerous studies showed older adult's slower and poorer picture naming ability compared to younger adults (Kent & Luszcz, 2002; Mitchell, 1989; Tsang & Lee, 2003). Studies on picture naming tasks showing naming deficits with aging also suggests the problem lies with access to phonological information of the word because subjects improved with phonological cueing (Au et al., 1995) and phonological cues leveled out the age differences (Thomas et al., 1977). Third, older people were observed using more pronouns than common nouns (Burke & MacKay, 1997), were more verbal during a TOT state (Brown & Nix, 1996), and produced more circumlocutions when word-searching during the BNT (Obler & Albert, 1985), all of which is likely because older people were less able to retrieve the proper word than the younger groups.

Possible Contradictory Findings

Although an increase in TOT probability was found for older adults (Burke et al., 1991), the speed or ability to resolve the TOT was equal to younger adults (Brown & Nix, 1996). The first part of this statement is consistent with the premise of an age decline in word-finding ability. The second part may appear to directly negate the proposed automatic and controlled processing effects on aging, where older people would be expected to be slower in resolving TOTs because TOT resolution would involve controlled processing. Further inspection of the findings provides clues to how these results could have materialized and how our hypotheses remain. First, the older group in Brown and Nix's (1996) experiment had significantly higher

verbal ability than the younger group. Higher verbal ability could result in better verbal search strategies, which would even out any differences between older and younger groups' controlled processing. More homogeneous groups may have shown poorer search strategies for older adults and hence, slower TOT resolution speeds, which would also support Hodgson and Ellis' (1998) findings that older adults respond to naming tasks slower than younger adults.

Summary

In summary, it appears that word-finding difficulties experienced by healthy older people does not indicate a problem with semantic aspects of the word, which is well preserved in aging, but to a deficit in the ability to retrieve the phonological aspects of speech production (see Chapter 4: Cognitive Models of Naming). The processing involved subsequent to the word-retrieval failure may also show an age-linked effect in a study design where the groups are more homogeneous and more representative of the population as a whole.

CHAPTER 7

MEASURING WORD FINDING

It is difficult to experimentally study word selection and production in speech, especially when investigating rare natural occurrences such as TOTs or normal or other word-finding problems (A.S. Brown, 1991). R. Brown and McNeill (1966), the first investigators of TOTs, set the initial framework for this type of study and an eclectic assortment of techniques has been used since. Although verbal fluency and discourse tests have been used clinically and experimentally to assess word-finding abilities, confrontational naming tasks, particularly picture naming, are most common (Gordon, 1997; Lezak, 2004; Nicholas et al., 1997). Studies (Dunn et al., 1989; Schmitter-Edgecombe et al., 2000) have compared picture-naming tests with alternative methods to assess word-finding skills and have drawn their own conclusions. Those who argue for discourse tests comment that discourse tests are more akin to spontaneous conversation and thus more appropriate to assess word-finding problems. Others argue that the neurocognitive process of naming a pictured object is similar to word production in spontaneous speech (Lezak, 2004; Loring, 1999) because both necessarily involve lexical access, a process Barry et al. (1997) refer to, somewhat eccentrically[4], as "lexicalization", which they define as:

> "the means by which a semantic or conceptual representation (e.g., <small mammal>, <domestic pet>, <can be trained to assist blind people>, <has a highly developed sense of

[4] In linguistics, the term "lexicalization" typically refers to the diachronic process by which common phrases come to be analyzed as a single word or lexical item. For example, presumably there was a time when the word White House was perceived and interpreted as a descriptive phrase as in "The new Presidential residence is that white house over there."

smell>, <barks>, etc.) is used to select the appropriate word, which then makes its phonological form ("dog") available." (p. 560)

Normal mapping between concept and lexical representation occurs rapidly and utilizes automatic cognitive processing.

Lexical access is fundamentally what neuropsychologists typically want to assess when they "measure word-finding." Unlike discourse tasks, picture naming test restricts a response to a single, concrete noun, reducing syntactic complexity to a bare minimum, but adding the process of object recognition as a prerequisite. Verbal fluency tests also remove complex syntactic processing by restricting responses and also without requiring the process of object recognition. However, verbal fluency tests do not elicit the specific semantic representation that is characteristic of subjects who report "word-finding" problems. It is the specific element of the searched for word during a TOT or word-finding complaint that is distressing. The current literature on three common methods used to assess word-finding problems will be discussed in the next sections followed by the rationale for selecting a picture naming test, the BNT, to construct a new method of measuring word-finding.

Picture Naming Tests

Naming faculties are commonly measured both clinically and experimentally with picture-naming tasks (Goodglass et al., 2001; Goulet et al., 1994; Lezak, 2004; Loring, 1999), also called confrontational naming. A highly researched instrument, the Boston Naming Test (BNT), is the most commonly used measure of word-finding (Lansing et al., 1999; Van Gorp, et al.; Welch et al., 1996), especially for research in a normal aging population (Schmitter-Edgecombe et al., 2000). The BNT was originally designed for one purpose, to detect aphasia in a clinical population (Goodglass et al., 2001), even though it is used in both clinical and research

settings (Mitrushina et al., 1999) and is considered to be helpful in identifying even mild word-finding problems (Thompson & Heaton, 1989). The BNT is a naming task where a person is presented with line drawings of objects ranging of high-frequency, high familiarity objects (e.g., tree) to those that are less frequent, less familiar (e.g., abacus) and is asked to name the picture. A prompting cue (semantic category) is given if the object's name is not perceived correctly. This is followed by a phonemic cue (first sound of the word) if the correct response is still not spontaneously produced. The task requires that a person visually interpret and identify the pictured object, mentally retrieve the correct word with its associated phonological representation and articulate the object's name, hence it is known as a "word-finding" task as well as a "naming" task.

Verbal Fluency Tests

Word fluency tests are productive naming tests that require that an individual "produce" in a restricted time period (typically one minute) as many words that begin with a specified letter of the alphabet, or to produce as many words as he or she can within functional semantic categories (e.g., foods, flowers). Previous studies have indicated a weak relationship between verbal fluency skills and word-finding abilities in healthy adults (Albert et al., 1988; Schmitter-Edgecombe et al., 2000) and moderate correlation between verbal fluency among clinical patients (Thompson & Heaton, 1989). Dunn et al. (1989) found a verbal fluency test (category animals) to be a more sensitive measure than a picture-naming task. Verbal fluency scores, unlike the picture-naming scores, were able to separate individuals without impairment from those with mild impairment in Dunn et al.'s study. In addition, the verbal fluency scores were helpful in distinguishing specific types of aphasia (e.g., fluent from nonfluent dysphasia) in this same study.

Discourse Tests

Discourse tests are naturalistic tests that measure a person's word-finding ability based on their ability to engage in discourse, or free-flowing conversation. Picture Description is the most common test format for measuring discourse (Brookshire, 1997). Other formats to elicit discourse for assessment include "Story Retelling" and "Interviews and Conversations." Schmitter-Edgecombe et al. (2000) discuss several benefits of using a discourse test over a picture-naming test to assess word-finding ability. First, discourse tests permit an individual to produce a more natural and spontaneous language sample that may more closely mimic the mode in which an individual experiences word-finding problems. Second, discourse tests allow for the identification of several types of word-finding errors (e.g., substitutions, empty words) that could offer clues for effective remediation. In contrast, naming tasks only require a one-word response which may less likely to occur in the context in which the subject's word-finding difficulties generally take place.

Conclusion

The choice of the BNT to measure of word finding in the present study has been made after careful consideration of all three methods. The first reason for this choice is tradition and familiarity. Word-finding difficulties have traditionally been assessed through visual object confrontation (Gordon, 1997; Goulet et al., 1994; Lopez et al, 2003; Nicholas et al., 1997) even though the language difficulties are frequently manifested in spontaneous speech (Loring, 1999). On the surface it may appear that word-finding in a confrontational task differs from those in spontaneous speech, but with respect to cognitive models of lexical access they do not; these two types of word-finding difficulty are "dissociable" and use similar neurocognitive processes (Lezak, 2004; Loring, 1999).

Verbal fluency tests were not selected because they lack the empirical support that naming tasks have received, and much ambiguity continues to exist between the relationship of age-related changes in verbal fluency and word-finding ability (Garcia & Orange, 1996). The less favorable results on picture naming tasks from Dunn et al. (1989) does not discredit the value of using a picture-naming test, such as the BNT, as a valid measure of word-finding ability in healthy adults. The Dunn et al. study had a relatively small sample size, using 22 dysphasic adults and 20 unimpaired adults, and the picture-naming test used was not an empirically driven test; it consisted of 15 of the "most frequently used animal names." Additionally, numerous studies have shown the influence of priming on word-finding abilities with both the positive (Burke & MacKay, 1997) and negative effects (Lambon Ralphe et al., 2002) that priming can have on naming performance. Contamination may have contributed to Dunn et al.'s findings because both the picture-naming test and the verbal fluency test used animals as the semantic category and the random ordering of the two tests could have caused a confound in the person's naming performance. Furthermore, Dunn et al. was trying to discriminate types of aphasia while the limits of picture naming on distinguishing between aphasic subtypes has been long recognized (original authors of the BNT, Kaplan, Goodglass, & Weintraub, 1983). Therefore, Dunn et al.'s findings are not generalizable to the BNT. In addition, word fluency tests are often thought to bear even less of a relationship to everyday speech than confrontational naming tasks (Brookshire, 1997).

Other reasons that the BNT was preferred over discourse tests and verbal fluency tests is its increased sensitivity, economy of time and ease of administration, and the wealth of literature available surrounding the instrument. Although discourse tests more closely imitate everyday experiences of word-finding problems, naming tasks may be a more sensitive instrument to

assess word-finding capabilities. Requiring a person to select one concrete word in a naming task, without the context of a narrative, may add complexity to the task because of the need for increased precision of selecting and retrieving a single "correct" word. This complexity of restriction also is absent in word fluency tests where one word is not sufficient. Often, people in a TOT state are very fluent and can produce verbal descriptions of the target word and generate other related words.

Compensatory mechanisms may not camouflage impairment with the BNT as easily as other tests. For example, older adults, or individuals with word-finding problems, may have learned to make up for word-retrieval problems by avoiding certain items (Schmitter-Edgecombe et al., 2000) that may go undetected in a discourse test. The subjective complaint about word-finding problems is "not being able to find the right word" (Sunderland et al., 1986) which is usually one specific word and may be unnoticed in either a verbal fluency or discourse test, but is acutely obvious to the individual (see Lovelace & Toweling, 1990, and Martin & Zimprich, 2003). Discourse tests, additionally, are time consuming and complex to administer and score. Responses generally are taken verbatim and protocols are often segmented into "T-units" (which are the smallest linguistic unit an utterance can be reduced to without leaving a fragment) before analyses. Currently no valid discourse test is available to clinically measure word-finding abilities in adults (Note: German (1991) developed a discourse test available to assess children). In contrast, the BNT is not complex or time-consuming to administer and score, it is frequently used as a measure of word-finding ability with much empirical research, and almost all versions of the BNT have very good reliability and validity (see Spreen & Strauss, 1998 for specific numbers). Unlike other tests, there is much support in BNT's usefulness in discriminating normal elderly individuals and those with dementia (Calero, Arnedo, Ruiz-Pedrosa, & Carnero,

2002; LaBarge et al., 1986; Lezak, 2004; Mack, Freed, Williams, & Henderson, 1992; Welch, Doineau, Johnson, & King, 1996) which is a common population in neuropsychological assessment. Picture naming tasks are also more conducive for studying word attributes (e.g., word frequency and length, age of acquisition) that affect naming ability (Barry et al., 1997) that are not possible with discourse tests or tests of verbal fluency. Most importantly, the word-finding functions found in regular conversation do not differ from word-finding functions in a confrontational naming task (Loring, 1999). Therefore, the wealth of available literature on the BNT combined with the convenience and the relatively straightforward means of examination makes the BNT the chosen instrument for the present project.

CHAPTER 8

BOSTON NAMING TEST (BNT)

The range of published articles using the BNT is extensive, and the number of studies providing normative data exemplifies its popularity (Lezak, 2004). The BNT has been investigated for both clinical and experimental purposes (see Feyereisen, 1997, and Goulet et al., 1994 for reviews of several BNT studies). Therefore, examination of individual studies must occur before drawing definitive conclusions or making generalizations. The results from many studies are mixed because the methods and aims of the studies vary in many respects--the version of the BNT utilized; whether they included age, education, gender, or intelligence as factors or variables; the age range and number of participants; and the method of administration and scoring–all of which must be considered prior to interpretation and comparative analyses. Before turning to our procedures for creating normative data on the BNT to create a new instrument, cautionary notes and a brief history of the BNT are necessary to provide a context and a rationale for the current investigation.

Caution with BNT Research

The BNT has been used in numerous studies to explore the efficiency of naming ability in various normal and clinical samples (Mitrushina et al., 1999). However, the studies vary in many respects. Several aspects of each study should be examined before any formative conclusion or generalization is made. First, the version of the BNT utilized is important. Currently there are many existing versions of the BNT: experimental 85-item version, 80-item version, standard 60-item version, as well as several shortened versions (Fastenau et al., 1998; Williams, Mack, & Henderson, 1989; See also Mitrushina et al., 1999 and Kent & Luszcz, 2002 for reviews of shortened versions) and versions for speakers of French, Spanish, Korean and Chinese (see Kim & Na, 1999; Roberts et al., 2002; Tsang & Lee., 2003, respectively). Many of

the shortened forms have been successfully validated for both normal controls and persons with dementia (Lansing et al., 1999) and with longitudinal data from a large sample (Kent & Luszcz, 2002). Second, the sample from which the normative data was derived must reflect the population being assessed for a measurement to be valid; otherwise there is a substantial risk of misdiagnosing naming impairment (Hawkins & Bender, 2002). Third, one must consider factors of age, education, intelligence or gender before clinically using test results, and these variables have not been thoroughly investigated in naming performances (Randolph et al., 1999). Fourth, administration procedures vary (Lopez et al., 2003), especially when determining a "failed" item (Ferman, Ivnick, & Lucas, 1998) or stimulus cue provisions (Mitrushina et al., 1999). The disagreement among neuropsychologists about administration approaches is so great that differing methods have produced significant differences in total score (see Lopez et al., 2003). Fifth, attention must be given to the aspect of performance that is reported (Mitrushina et al., 1999). Some studies report the percentage of correct responses per item, others report total score or scaled score, and some report error analyses with different error classification systems. Finally, different age intervals are used by different studies. Some studies primarily use decade age intervals and others use shorter age intervals. Smaller age intervals between comparison groups may conceal potential age effects (Au et al., 1995). With all of the variations between the studies on the BNT, it was no surprise there was mixed results. However, the plethora of published information concerning the BNT can be very useful after all of the aforementioned perspectives are carefully considered.

Some of the cautions will be addressed in the next few sections. Two of the cautions listed warranted their own chapters: Chapter 9: Individual Variables and BNT Performances, and Chapter 10: BNT Administration and Scoring.

History of BNT Versions

In the most current BNT test manual, Dr. Harold Goodglass states: "The obvious method of testing patients for word finding difficulty is to present pictures or questions requiring the selection of a particular word in response" (Goodglass et al., 2001, p. 7). In 1960, Dr. Harold Goodglass received a grant from the National Institute of Health to test people with the original form of the Boston Diagnostic Aphasia Examination (BDEA) for aphasia. The first version of the BNT was published in 1978 by Kaplan, Goodglass, and Weintraub. This original version of the BNT (Kaplan et al., 1978) was considered an "experimental version" consisting of 85 line drawings intended to supplement the BDAE. In 1983, a modified version was published that included 60 of the original 85 drawings, arranged in order of increasing difficulty, and was still considered a supplement to, rather than a part of, the BDEA (Kaplan, Goodglass, & Weintraub, 1983). The most current, third edition of the BNT, published in 2001 (Goodglass, Kaplan, & Barresi, 2001), incorporates the BNT in the new version of the BDEA which helps examiners determine the extent to which aphasic individuals can recognize the pictures that they are unable to name.

In addition to being used as a stand-alone test, or as part of the BDEA, the BNT has also been part of several neuropsychological batteries (see also Lezak, 2004 for a descriptions of each): Halstead Russell Neuropsychological Evaluation System (HRNES) (Russell & Starkey, 1993); California Neuropsychological Screening Battery-Revised (CNS-R) (Bowler, Thaler, Law, & Becker, 1990); and, using the Spanish version of the BNT, Neuropsychological Screening Battery for Hispanics (NeSBHIS) (Pontón, Satz, Herrera, et al., 1996).

Existing Normative Data

A great deal of normative data is available for the BNT; only a few select norms will be presented. Mitrushina et al. (1999) published a comprehensive review of 19 norm sets, many of which are not presented here. Again, the large number of studies contributing normative data for the BNT is considered further testament to its popularity (Lezak, 2004). Normative data for the 85-item experimental edition of the BNT (Kaplan, et al., 1978) was first published by Borod, Goodglass, and Kaplan (1980) using 147 normal males, grouped into five age categories (25-39, 40-49, 50-59, 60-69, and 70-85). More norms for this version have been published subsequently for individuals at different age levels (LaBarge et al., 1986; Nicholas et al., 1985).

The second edition of the BNT (Kaplan, et al., 1983) provided normative data on the 60-item version for 84 normal adults, aged 18 through 59 years of age, broken down into two educational groups and five age groups; and for 82 aphasic patients grouped by aphasia severity level. Heaton et al. (1991) and Thompson and Heaton (1989) found high correlations between the 85-item experimental version and the 60-item versions of the BNT. Van Gorp et al. (1986) subsequently published normative data on this edition for 78 normal adults, extending the age perimeters to include 59 to 95 year olds. However, the Van Gorp et al. normative data has been scrutinized because of its "superhuman" population which included only very high-functioning older adults (e.g., Mean Full-Scale IQ = 122).

The country of origin of members of both the normative sample and corresponding population being assessed must be considered before using published BNT norms (Kent & Luszcz, 2002). Different versions of the BNT with accompanying norms are available for diverse populations. Korean (Kim & Na, 1999), Australian (Worrall et al., 1995), and Chinese (see

Tsang & Lee, 2003) norms are available, and Ross & Lichtenberg (1997) offer norms for an American, elderly, urban medical sample.

Shortened Versions

Several shortened versions of the BNT are available to offer a more streamlined measure, to reduce the demands on severely impaired or elderly patients, and for test-retest purposes (for example, Fastenau et al., 1998; Lansing, et al., 1999; Mack et al., 1992; Williams, Mack, & Henderson, 1989; see Kent & Luszcz, 2002 for review and a table of norms for several shortened versions). Mitrushina et al. (1998) presented reviews for many of the available shortened forms at the time of their publication. Goodglass, Kaplan & Barresi (2001), the authors of the most recent version of the BNT, present restandardized normative data from their previous norms and offer a 15-item short form of the BNT that is bound in the beginning of the BNT stimulus booklet as well as an updated standardization of normative data derived from 85 aphasic subjects and 15 normal elderly volunteers from the community. Fastenau et al. (1998) present normative data for four existing 15-item and two 30-item shortened versions for the Boston Naming Test that were validated using 108 healthy adults, ages 57-85. Kent and Luszcz (2002) judged the shortened versions' normative data to be "inadequate" (p. 561) for assessing naming ability over time and produced normative data from longitudinal data for four shortened versions of the BNT with a large population of community-dwelling Australians.

Conclusion

In general, cautious interpretation is necessary when assessing any individual that is not adequately represented in the normative data on the BNT version used, including demographical region, level of education, verbal ability, and ethnicity. Sample size is not the problem; the issue is representativeness (Hawkins & Bender, 2002).

CHAPTER 9

INDIVIDUAL VARIABLES AND BNT PERFORMANCES

Demographic Factors and BNT

The general effects of age, gender, education and IQ, and verbal skills on cognitive ability, and in naming in particular, have already been discussed in an earlier chapter (in Chapter 3: Factors affecting Naming Ability). With specific regard to naming and the BNT, the effects of age, education, and gender on BNT performance have not been addressed consistently in the literature (see Randolph, Lansing, et al., 1999, for a more complete review of these issues). The following sections contain research findings on the above variables and how they relate specifically to BNT performances.

Age

The results concerning an age-related decline in BNT naming scores are mixed (Feyereisen, 1997; Goulet et al., 1994). Several reasons for the discrepancy have already been discussed. Mixed results specific to the BNT are methodological issues, subject selection criteria, age ranges and age groupings, sample sizes, and variations in administration and scoring the BNT. In addition to the "cautions" regarding BNT research listed above, cohort effects can help explain some of the divergent findings with age and the BNT.

Cohort Effects

Cross-sectional research on naming abilities can produce cohort effects, which reports age-related differences and not age-related changes. Cohort effects could affect BNT scores because basic differences between young and old individuals may be exaggerated in cross-sectional designs. Old and young people differ in respect to verbal knowledge acquisition (Randolph et al., 1999) and certain items on the BNT have been found be more discriminative of semantic knowledge than naming differences between age groups (Schmitter-Edgecombe et al.,

2000). Schmitter-Edgecombe et al. recommend close examination of the psychometric properties of individual BNT items in order to identify confounds related to cohort issues and to IQ and education. The findings of Schmitter-Edgecombe and colleagues showed a significant cohort effect on four BNT items – yoke, trellis, palette, and abacus – that were missed by 72% of their younger participants, but less than 32% of the older participants. Phonemic cues did not help the younger participants suggesting it was not a problem of retrieval but one of semantic knowledge. Longitudinal research designs could eliminate possible cohort effects and underscore some of the true changes of naming ability with increasing age.

Longitudinal Studies

A longitudinal study by Cruice et al. (2002) illustrates the potential for cohort effects in BNT research. Two separate analyses of Cruice et al.'s data indicated a weak, but significant age-related decline when analyzed as a cross sectional study, but no significant age-related decline was evident in the longitudinal design. Unfortunately, longitudinal data for the BNT is restricted (Kent & Luszcz, 2002). One study of healthy groups of adults (Au et al., 1985) found that all groups except those in the 30-year group showed a decline in BNT naming scores over a 7 year span, suggesting cohort effects were not the reason for the age differences found in BNT scores, and that decline in naming ability is a phenomenon which occurs in natural aging. Further analysis in Au et al. study showed those in the 70-year group found cues to be less helpful over time, possibly indicating more difficulty in word retrieval due to decreased processing efficiency in elderly participants. Findings from a more recent longitudinal study by Hickman et al. (2000) showed a practice effect in younger participants that was not apparent in the older participants. Combining interpretations from Au et al. and Hickman et al., one could state that older people's slower processing speed hinders their ability to learn from practice. This supports the hypothesis

of decreased controlled processing abilities with age that will show in increased BNT-L (our new measure) latency responses with age.

Like cross-sectional studies, longitudinal studies also produced mixed results on naming and aging. Kent and Luszcz (2002) produced normative data and examined naming ability at 2 years (*N*=803) and at 6 years (*N*=326) (8 years total) in a large sample of community-dwelling Australians and found true naming declines in BNT performances occurring between ages 80-84 years. On the other hand, other longitudinal studies, one on an Australian population (Cruice et al., 2000) and another on an American population (Hickman et al., 2000) found no age-related effects on BNT naming ability over a four-year period. These latter findings suggest that naming performances remain relatively stable with aging.

Variability

There may be doubt about a natural age-related decline in naming ability (MacKay et al., 2005; Goulet et al., 1994), but there is little doubt that more variance exists in BNT performances of older people (Nicholas et al., 1989; Van Gorp et al., 1986; Welch et al., 1996). The range of scores expands and standard deviations become larger when the age groups increase (Nicholas et al., 1989; Van Gorp et al.) and when education level decreases (Borod et al., 1980; Hawkins et al., 1993; Nicholas et al., 1985; Thompson & Heaton, 1989; Van Gorp et al.; Welch et al.; Worrall et al., 1995). The variability has often been attributed to the heterogeneity in elder populations with uneven demographics and unequal health statuses (Cruice et al., 2000; Hickman et al., 2000; Randolph et al., 1999). The variability, especially within an older cohort, prompted one author to declare the current use of the BNT alone to be inadequate "to determine pathological naming difficulties when there is such a wide rage of 'normal' scores" within this population (Cruice et al., 2000, p. 151).

Education and Intelligence

Several studies have found significant relationships between level of education and BNT scores (Borod et al., 1980; Hawkins et al., 1993; Heaton et al., 1999; Henderson et al., 1998; Tombaugh & Hubley, 1998; Kent & Luszcz, 2002; Kim & Na, 1999; Lansing et al., 1999; Nicholas et al., 1989; Ross et al., 1995; Spreen & Strauss, 1998; Thompson & Heaton, 1989; Welch et al., 1996; Whitfield et al., 2000; Worrall et al., 1995). Some multiple regression analyses (MRA) have shown years of education to be the best predictor of BNT scores (Tombaugh & Hubley) and yet other MRAs showed no effect of education on predicting BNT scores (Fastenau et al., 1998; Kent & Luszcz). The interaction of age and education had significant effects on BNT scores in many studies (Borod et al.; Kim & Na; Randolph et al., 1999; Welch et al., 1996), with age and education as the best predictors of BNT scores (Welch et al.). Furthermore, the relationship between education and BNT scores are reported throughout the educational range, including higher educational levels (Hawkins et al., 1993; Heaton et al., 1999).

Years ago the writers of the BNT (Goodglass & Kaplan, 1983) addressed the issue of different educational levels by including two sets of normative data, one for 12 years or less of education, and one with more than 12 years of education. However, the sample size for the lower education norms was small (N=15). As a result, using the published BNT norms was found to produce high false positive rates in subjects with lower education or intelligence. Hawkins et al. (1993) studied the BNT performance of normal participants and discovered that those with less than the equivalent of a twelfth-grade vocabulary level scored below the published norms in the 1983 manual authored by Kaplan, Goodglass, & Weintraub. When creating BNT normative data it is important to ascertain, to the greatest possible extent, that the normative sample is

representative of the population, or that the effects of education and/or IQ are adjusted before using the comparative data (Hawkins & Bender, 2002). Subsequent normative data has been criticized because of the higher than average education and intellect of the published subject pool (e.g., Welch et al., 1986).

Vocabulary

The studies that have directly related BNT performance to vocabulary have found significant correlations (Albert et al., 1989; Hawkins et al., 1993; Killgore & Adams, 1998; Thompson & Heaton, 1989). Using the WAIS-R Vocabulary scores, Albert et al., (1989), Killgore and Adams (1998), and Thompson and Heaton (1989) all found a significant relationship between vocabulary scores and BNT scores. Killgore and Adams created BNT cutoff scores based on obtained vocabulary scores. The only other study exploring vocabulary and BNT performances used a reading vocabulary test (Hawkins, et al.).

Reading vocabulary tests could provide indicators of one's verbal ability. Tests of reading vocabulary/recognition have been suggested as a guide for BNT performance expectations (Hawkins et al., 1993). Reading vocabulary/recognition tests also signify premorbid abilities (McGurn et al., 2004) and are resistant to age-related cognitive declines (Tombaugh, 1996) and disease (Lezak, 2004). Using Level 7-9 Gates-MacGinite Reading Vocabulary Test (G-MRVT), Hawkins et al. found reading vocabulary to be highly correlated with BNT scores and offers guidelines to complement BNT norms based on G-MRVT scores. For those not familiar with the G-MRVT, it is a test that "involves simple word recognition" (Lezak, p. 523).

The studies that used WAIS-R Vocabulary scores reported very similar correlations between WAIS-R Vocabulary raw score and BNT performance: Tombaugh and Hubley (1997) reported $r = .53$; Killgore and Adams (1999) reported $r = .65$; and Thompson and Heaton (1998)

reported $r = .79$. These are similar to the Hawkins et al. (1983) finding ($r = .81$) between the G-MRVT and BNT scores. These correlations have implications regarding how the BNT-L norms will be created.

Education vs. Verbal IQ for BNT-L Norm Stratification

If necessary, BNT-L scores will use the WTAR to obtain estimated Verbal IQ for stratifying normative data, first, because BNT scores typically correlate more strongly with vocabulary (see Albert et al., 1989; Killgore & Adams, 1998; Thompson & Heaton, 1989) than with education (see for comparison: Borod et al., 1980; Heaton et al., 1999; Henderson et al., 1998; Tombaugh & Hubley, 1998; Kent & Luszcz, 2002; Kim & Na, 1999; Lansing et al., 1999; Nicholas et al., 1989; Ross et al., 1995; Spreen & Strauss, 1998; Welch et al., 1996; Whitfield et al., 2000; Worrall et al., 1995), and second, because similar tests to the WTAR (e.g., NART, G-MRVT) have shown resistance to brain compromise (Lezak, 2004; McGurn et al., 2004), are relatively quick and easy to administer and score (Tombaugh, 1996), and have been established high correlates with standard the BNT (Hawkins et al., 1993). Finally, although the number of years of education is a static demographic, variability within a given educational stratum can provide misleading norms (e.g., norms published by Welch et al., 1986).

Explanation for Mixed Findings on Education and Vocabulary

Despite the high correlations found in the studies listed above, other studies have not found a significant relationship between education and BNT scores (Albert et al., 1988; Cruice et al., 2002; Farmer, 1991; Fastenau et al., 1998; LaBarge et al., 1986; Nicholas et al., 1985). Two explanations why some studies have found a sizeable relationship between education and BNT scores and others have not involve the BNT's psychometric properties and the normative data used (Hawkins & Bender, 2002).

Psychometric Properties of BNT

BNT scores for normal subjects do not fit a normal distribution (Mitrushina et al., 1999). BNT scores are skewed toward the high end, and most scores cluster around the mean (Hamby, Bardi, & Wilkins, 1997). This negative skew (asymmetry) and extreme kurtosis (peakedness) causes the scores to have a slender, high peak off to the right. Depending on the sample composition, a short tail will be present if the sample comprises subjects with higher education and stronger vocabularies. Distributions such as this reflect an insensitivity of the test at levels of average abilities and above (Hamby et al.). When normative samples lack satisfactory representation of subjects with lower abilities, an immense limitation on score variability will limit "the magnitude of correlation coefficients that can be found with other variables such as education or vocabulary" (Hawkins & Bender, p. 1148; see also Fastenau et al., 1998).

A consequence of BNT psychometric properties is that the BNT can be considered a pass/fail test where normal subjects' scores cluster close to the mean (Hamby et al., 1997; Hawkins & Bender, 2002). This distribution masks the reality that subjects with limited vocabularies score significantly below subjects with average, or above average scores. In addition, distributions with extreme kurtosis yield small standard deviations, in which a small deviation can register as an abnormal score without adequate corrections for education. This can explain the findings of Hawkins et al. (1993) with the use of norms obtained from an unmatched normative sample. Furthermore, the distribution described above can explain the variability reported in BNT scores and decreased education in the sample (Borod et al., 1980; Hawkins et al., 1993; Nicholas et al., 1985; Thompson & Heaton, 1989; Van Gorp et al., 1986; Welch et al., 1996; Worrall et al., 1995).

Normative Sample

The second explanation for the discrepancies in education effects and BNT scores is the restricted demographic that comprises many normative samples (Hawkins & Bender, 2002). Randolph et al. (1999) cautioned readers that the normative data produced was only applicable to subjects with a demographic similar to that of their sample. His sample, like most of the samples comprising BNT normative data, was not representative of the US population. Most of the normative data published for the BNT contains samples with 12 years of education or more, which does not represent a significant portion of the US population. The US Bureau of Census population data as reported by the Psychological Corporation, 1997, shows that 30% of the US population aged 65-69 has less than 12 years of education, and over 66% have 12 years of education or less. For younger segments aged 20-44, 46% have 12 years of education or less, and 11% have 11 years or less. As a result, much of the current normative data may result in misleading norms, which can be exacerbated with elderly populations (Cruice et al., 2000; Hawkins & Bender, 2002; Hickman et al., 2000; Randolph et al., 1999).

Regression-based corrections for demographics have been suggested, and criticized, in the literature (see Fastenau, 1998; Heaton et al., 1999). The sufficiency of these guidelines also depends on the representativeness of the population from which they were derived (for elaboration, see Hawkins & Bender, 2002).

Gender

Gender effects on BNT scores are equivocal. Although several studies have found no effects of BNT scores and gender (Azrin et al., 1996; Barker-Collo, 2001; Cruice et al., 2000; Fastenau et al., 1998; Henderson et al., 1998; Kent & Luszcz, 2002; Kim & Na, 1999; LaBarge et al., 1996), other studies have found a gender difference in BNT performance, with males

scoring significantly higher (Lansing et al., 1999; Randolph et al., 1999; Tombaugh & Hubley, 1997; Welch et al., 1996), up to 4 points on average in one study (Welch et al.). This finding was surprising given the traditional theory that females are superior on verbal tasks. Further inspection of results reported by Randolph et al. determined that the gender differences were specific to individual items that were more salient for each gender, with nearly four times more items more salient for males (see article for a listing). Welch et al. found women only scored higher than men on two items: *asparagus* and *palette* (these two were also on Randolph et al.'s list for higher scores for women). Welch attributed the difference in scores to male's occupations and male propensity to use terms such as *compass*, *protractor*, or *yoke.* It is probable that the male advantage observed on BNT performances is not strictly based in language or memory, particularly since the results in Randolph et al.'s study appear to have been entirely stimulus driven (not in Welch et al.'s sample, the men had slightly more education). However, if gender effects were due to stimulus factors alone, it is unclear why several studies did not produce a gender effect while using the same stimuli; and actually, most studies found no gender effects. The overall mixed results of gender effects at all (Randolph et al., 1999; Spreen & Strauss, 1998; Welch et al., 1986) make gender a weak variable in BNT performance (Lezak, 2004).

Race and Culture

There is some evidence that the performance of minority populations may be poorer on confrontational naming tests (Randolph et al., 1999). In fact, multiple regression analyses produced separate prediction equations for African Americans and European Americans for BNT performances but not on other cognitive performance measures (Whitfield et al., 2000). African American participants in a study by Azrin et al. (1996) produced more alternative names for BNT items (e.g., walkers for *stilts*) that resulted in overall poorer scores. The Azrin et al. study

also determined that certain names given to items on the BNT could be regionally biased. For example, in Southern Mississippi a regionally correct word for "harmonica" is "mouthorgan." It should be noted that after this publication the most recent version of the BNT (Goodglass et al., 2001) permits "mouth organ" as an acceptable response for *harmonica*.

Different norms and modifications of the BNT test items may also be needed for different ethnicities (Barker-Collo, 2001; Kim & Na, 1999; Tsang & Lee, 2003) due to the vocabulary demands (Calero et al., 2002) or interference from bilingual capabilities (Roberts et al., 2002). In these cases, poor BNT scores may not be due to anomia but to lexical ignorance of the name of the item presented (Calero et al.) or to responses that are culturally or regionally appropriate but are not "correct" for BNT responses (Azrin et al., 1996; Cruice et al., 2000). Several authors have addressed this issue. To compare levels of item difficulty on geographical region, Barker-Collo (2001) provided a "difficulty index" for the 60-item BNT for populations from Canada, Australia, and New Zealand. An additional attempt to address cultural issues, Roberts et al. (2002) provided suggestions for a more "lenient" scoring system for 18 BNT items that included alternate responses and was less ethnically biased.

Conclusion

It is clear that factors of age, education, verbal IQ, gender, and culture must be considered when interpreting the data on BNT scores, and when constructing a new instrument. Up to the date of publication, Lansing et al. (1999) examined all versions of the BNT and found a significant effect of age, education, and gender on all previously published forms. In addition, a strong bias is apparent for specific BNT test items for gender and race which demands great care when analyzing items for inclusion in the BNT-L. The WTAR will be administered along with the BNT-L to all participants to provide the best starting point for BNT performance

expectations. Fervent attempts were made in the recruitment of a normative sample for the BNT-L that is representative of the current US population, or at best, representative of the population in most clinical settings to offer valid normative expectations for participants with lower education or more restricted vocabularies.

CHAPTER 10

BNT RESEARCH AND FACTORS TO CONSIDER WHEN SELECTING ITEMS FOR BNT-L

BNT Administration and Scoring

Problems have been reported with the administration and scoring of the BNT. The Kaplan et al. (1983) administration and scoring instructions have been found not comprehensive enough to ensure consistency in the way examiners administer the test and score individual items (Brookshire, 1997; Ferman, et al., 1998; Lopez et al., 2003; Nicholas et al., 1989). Lopez et al. compared three different interpretations of scoring methods used by practicing clinicians, all three interpretations seemed correct but each produced varied results. The lack of consistency in administration and scoring also causes concern when using the published norms. Nicholas et al. offer revised and expanded instructions for administration, response coding, scoring and prompting procedures to improve interexaminer and intraexaminer reliability. Only when administration and scoring methods in clinical assessment match the administration and scoring procedures that were used during acquisition of normative data will a measure be valid.

Item Analysis and Item Selection

A few stimulus items on the BNT are considered "ambiguous or visually confusing" (Nicholas et al., 1989, p. 570) and must be examined in order to preserve construct validity of the BNT-L.

Alternative Responses

Alternative responses, or synonyms of the target word are common to healthy American (Azrin et al., 1996; Hubley and Tombaugh, 1998; Nicholas et al., 1989), Australian (Cruice et al., 2002), French and Spanish (Roberts et al., 2002) populations. Nicholas, Brookshire et al. (1989) found that 80% of incorrect responses from normal American adults fell into two

categories: Related Name and Don't Know. Visual misperception or not paying attention to details of the pictures seemed to cause responses in the Related Name category (e.g., saying schoolhouse for *house* or hippopotamus for *rhinoceros*). Fifty percent of the "incorrect responses" in Nicholas and colleagues' study fit in the Related Name category, subjects provided a nearly correct response but not the word provided in the manual. Hubley and Tombaugh (1998) examined error distributions and found that similar semantic errors were given to specific items (e.g., lock for *latch*). Similarly, Goldstein et al. (2000, as cited by Lezak, 2004) listed frequent "alternative responses" received on BNT items – *mask* (false face, face); *pretzel* (snake, worm), *harmonica* (harp, mouth organ), *stilts* (tommy walkers, walking sticks). The frequency of alternate, or related, responses in Goldstein's study were associated with age, education, race, and geographic region. All subjects were nonimpaired and these "incorrect responses" could erroneously suggest a problem with word finding when the error type was more likely due to confusing items or to careless examination of the test item (Nicholas et al., 1989). The most recent published BNT Scoring Booklet accounts for this research and now permits false face for *mask* and harp or mouth organ for *harmonica*.

Name Agreement

"Name agreement" has been found to be an independent predictor of naming accuracy in a picture-naming task (Hodgson & Ellis, 1998) (see discussion in Chapter 5: Response Latencies). BNT items that produce a "Related Name" response may possess low name agreement, or have more than one potential answer, which has been shown to increase response times on naming tasks (Vitkovitch & Tyrrell, 1995). "Related Name" responses could also provide an explanation for types of responses typically coded as errors on the standard BNT format. For example, a response of a culturally "correct Related Name" (e.g., tom(my) walkers -> *stilts*) reflects post-

semantic pre-phonological processing and most likely reflects individual dialectal differences. On the other hand, an "incorrect Related Name" response (e.g., horse -> *unicorn*; tripod -> *easel*) likely reflects an error at the level of semantic encoding and potentially indicates a word-finding problem. It could be that the latter type of response illustrates either what Nicholas et al. (1989) described as "[normal] subjects responding before they looked carefully at the details of the drawing" (p. 576) or a case where true semantic problems exist. The present investigation of response styles to BNT items may identify items with low name agreement that must be considered when selecting items and determining scoring rules for the new testing instrument. Obtaining qualitative information pertaining to individual items will occur during data collection for the normative sample.

Speed-Accuracy Tradeoff

Some existing items on the original BNT may be inappropriate for the modified BNT test that includes latency times. Latency times in themselves may affect the inclusion of certain test items. Research has shown that an individual's performance can vary depending on the subject's awareness of being timed (Duncan, 1986). The large number of errors found in Nicholas et al.'s (1989) healthy sample was attributed to hasty responses and careless examination of the specific test items. This function has been described as a speed-accuracy tradeoff (Duncan, 1986) that is most common in timed tasks–an individual sacrifices accuracy for speed, or vice versa. For example, one may be careful and produce the correct response but at the cost of using additional time, or on the other hand, one may sacrifice being correct because they are trying to be quick. Since this response style was observed even when the task was not timed (Nicholas et al., 1989), administration procedures for the BNT-L will contain clear and specific instructions for participants to respond *both* quickly and accurately.

Summary

In addition to age, education, and gender effects (Hubley & Tombaugh, 1998), regional and cultural differences must be taken into account with selecting individual test items for the BNT-L. Studies have shown certain BNT items elicit alternative responses in unimpaired individuals of differing race (Azrin, et al., 1996), ethnicity (Calero et al., 2002; Roberts et al., 2002) or region (Cruice et al., 2002; Goldstein et al., 2000). These factors could feasibly result in a false positive indicator for naming impairment. Furthermore, discrepancies in missed items between older and younger healthy adults as found by Schmitter-Edgecombe et al. (2000) and differences in response accuracy between males and females (e.g. Randolph et al., 1999; Welch et al., 1989) will be analyzed to assess the appropriateness for inclusion of items in the completed instrument intended for a wide age range. In addition to careful administrative considerations, individual items must be inspected in terms of age (Schmitter-Edgecombe et al., 2000), cultural (Roberts et al., 2002), educational (Hawkins et al., 1993), and gender (Randolph et al., 1999; Welch et al., 1986) influences when creating the BNT-L to prevent scoring items incorrectly and misattributing word-finding impairment when there is none. In particular, individual BNT items highlighted in the literature will be examined and permissible variants, or alternate responses, will be considered in conjunction with subjective and qualitative data collected from our normative population before selecting items for the shortened instrument. Due to the inherent low name agreement, and normal participant's frequency of responding with a "related name" (e.g., mouth organ/*harmonica,* lock/*latch*), it seems reasonable to accept selected alternative response to select BNT-L target words.

Error Types on BNT with Age

Much qualitative research followed Borod et al.'s (1980) publication pointing to problems with word retrieval in normal aging. Researchers looked at the qualitative differences in error responses in attempt to understand the disparities between younger and older individual naming abilities (e.g., Albert et al., 1988; Nicholas et al., 1985). Classification of naming errors is based on the presumption that error types reflect underlying mechanisms of word retrieval (Mitrushina et al., 1999). For example, identifying individual error types may reveal variations in younger and older individual's word-search strategies or weaknesses at specific levels of processing error, such as at a perceptual level, semantic level or phonemic level.

It appears that healthy subjects of all ages produce similar types of error responses despite a few significant age differences for some error types (e.g. older people produce more verbalizations relating to the target word, see Brown & Nix, 1996; Burke & MacKay, 1997; Nicholas et al., 1985; Obler & Albert, 1985). Several error analysis systems for the BNT have since been published: see Mitrushina et al. (1999) for descriptions and tables of several error classification systems; and, Kohn and Goodglass (1985) or Goodglass et al. (2001) for an error analysis system written by the original author of the BNT. Goodglass, Wingfield, and Hyde (1998) published a corpus of naming errors on the BNT for an aphasic population in attempt to establish error patterns to draw inferences. An important finding in their analysis was that normal participants generated similar errors and self-corrections as the aphasic participants, supporting the notion presented earlier that problems with word finding lies along a continuum.

Conclusion on BNT Research

Inconsistent results are common in scientific research. Meta-analytical techniques have been used to help clarify some of the inconsistencies found in BNT research. A quantitative

combination of research results from 32 BNT studies on aging was completed by Feyereisen (1997) and different conclusions were generated from Goulet et al.'s (1994) review of 25 BNT studies: Whereas Goulet et al. concluded no strong evidence supports an age-related decline in naming performance, Feyereisen concluded age effects do occur in naming accuracy, and it most likely begins after age 70. According to another review by Kent and Luszcz's (2003), age-declines occur after age 80. Feyereisen discovered the "inconsistencies" found in studies showing older people performed better on naming tasks than younger people occurred when the young were compared to an intermediate group, and not to the oldest group, if an oldest group was even present. In fact, in two out of three studies that reported no age-related decrease in BNT scores had no participants older than age 70, and below age 70 years is not that "old" by today's standards.

CHAPTER 11

TEST CONSTRUCTION: BNT LATENCY TEST (BNT-L)

No formal measure of word finding is available, however, no formal test construction is needed. Modification of an already existing test, the BNT, can potentially increase the utility of this instrument for better use in a normal population, or individuals who are bordering impairment. The latency to respond to items on the BNT, measured in whole seconds, will be the method used to assess naming ability is in this study. The name of this revised instrument will be called the Boston Naming Test of Latencies (BNT-L). We will use a standard digital stopwatch to record latencies for the BNT-L even though most response latencies in the literature were recorded by a computer program (e.g., Dunn et al., 1989; Feyereisen, 1998; Stern et al., 1991). This macro measurement of response time will be feasible for clinicians to use in the office and has been sufficient to distinguish normal from abnormal naming ability (Brookshire, 1971) and age differences using the BNT (Tsang & Lee, 2003). The BNT was selected because it is a well-known test commonly used to assess the naming abilities (Lansing et al., 1999; Lezak, 2004; Spreen & Strauss, 1998; Welch et al., 1996) and it is the most commonly used instrument by clinicians to formally assess naming ability (Van Gorp et al., 1989).

Concepts and elements of test construction must be considered to ensure the appropriateness of BNT latencies as our testing instrument. Most neuropsychological measures are developed either for use with an impaired population or with a normal population (Christensen, Multhaup, Nordstrom, & Voss, 1991). Measures intended for a normal population are helpful in recognizing deficits compared to an expected level of performance, but they can display floor effects that make them insensitive to differences in severity of impaired populations. On the other hand, measures intended for impaired populations may provide the

appropriate index of severity, but they reveal ceiling effects which make them insensitive to differences in normal and mildly impaired populations.

The BNT was specifically constructed as a measurement for an impaired population (Kaplan et al., 1978) and it is well known that BNT scores are not normally distributed (Cruice et al., 2002; Hamby et al., 1997; Hawkins & Bender, 2002; Mitrushina et al., 1999). Therefore, before using the BNT as an instrument to separate normal from abnormal naming abilities, careful consideration must be given to the likelihood of ceiling effects, and to the potential of the instrument to being insensitive to the very population we want to measure. Item difficulty is the first consideration regarding ceiling effects. An ideal level of item difficulty is .90 (Christensen et al., 1991), which means that 90% of a given sample will respond correctly to an individual item. We expect our prospective sample will provide 90% accuracy for each BNT item administered after careful analyses and careful selection of individual items to include in the BNT-L. In regards to test sensitivity, we believe the latency times of the responses will differentiate healthy normal responses from abnormal responses that may indicate pathology. Latency to respond provides much more sensitivity than accuracy responses within the 20 second administration rules. As stated previously, normal, uninhibited word production is typically considered an automatic process that occurs in less than 1s. There is much lexical difference in a correct response at 1.5s compared to 18s, yet each response could be scored equally on the standard BNT.

In general, the number of test items at an appropriate level of difficulty for both cognitively impaired and healthy populations is usually small for most scales (Christensen et al., 1991). Performance disparities are typically found in instruments that measure both normal and impaired populations. Response latencies and appropriate normative data should lessen these

disparities. Accurate performance estimates for impaired individuals requires tests that are constructed with data from both impaired and normal groups. We are using an already constructed test, however, we will collect data from both impaired and normal groups with collecting normative data. First, a moderate size sample of normal, healthy adults using BNT response latencies will be used to create the BNT-L, followed by measurements of impaired samples of individuals -- with diagnosed mild-to-moderate head trauma and a dementia -- that can determine the validity and utility of the BNT-L.

We hypothesize that most normal, healthy adults will be able to respond "automatically" to the BNT items within one or two measurable seconds, and responses outside this parameter suggests "controlled processing" (see Stern et al., 1991) which utilizes other cognitive functions that could indicate a bona fide "word-finding" problem. Once a person fails to retrieve the preferred word, extra-processing, or "controlled processing" is involved. Nearly every speaking person has experienced a word-finding problem or a TOT state, and concerns about this experience increase with age (Lovelace and Twohig, 1990). The goal of this experiment is to norm the BNT using response latencies to determine if latency response times will provide a more sensitive indicator of natural age-related declines as well as providing a measure to separate normal from abnormal naming abilities.

New Normative Data

Mitrushina et al. (1999) provided a compendium of neuropsychological test norms for several common instruments, including the BNT. Of all the normative studies available for the BNT, certain criterion had to meet for inclusion in their summary. Investigations were reviewed if the study used a large, well-defined sample (at least 50); the 60-item version was used; the "total score" represented the total number of correct responses, with or without a stimulus cue;

and groupings for age intervals were limited and expressly stated. An adequate description of sample composition was also important, such as: exclusion criteria for medical and psychiatric problems; educational level; and overall intellectual level of the sample. At minimum, group mean and standard deviation had to be presented. Judging Mitrushina et al.'s standards to be sound, our study will endeavor to achieve or surpass these guidelines.

Creating Normative Data for the BNT-L:

Analyzing Latency Times with Demographic Variables

Several issues of consideration are present with the proposed project after extensive review the literature. First, latency data is not normally distributed, and is positively skewed (Duncan, 1986) which eliminates the assumption of normality essential to many statistical methods. Second, gender, age, education and IQ have been found to be frequent predictors of overall BNT performance (see previous chapter, "Variables that affect naming") and must be factored into or removed from prospective normative data. Third, BNT items are not equally weighted prohibiting classical item analysis. Fourth, a final outcome is to derive an instrument with clinical utility and ease of scoring and administration. Neuropsychologists will not use an instrument that is too time consuming or requires too many statistical computations.

In order to appease all of the above considerations as well as maintain the variability in the predictors (demographics), nonparametric statistical methods for item response theory (IRT) will be used. IRT's general objective is to construct reliable and valid and good estimates of abilities of individual respondents.

CHAPTER 12

THREATS TO VALIDITY

Recruitment Bias

Recruitment bias is a potential threat to validity. Healthy volunteers for psychological research tend to have higher educational levels and could possibly possess vocabularies that are at a higher level relative to a clinical population reporting equal educational levels (Hawkins & Bender, 2002). There may also be much variability within a given educational stratum. For example, Van Gorp et al. (1986) published BNT normative data for a sample with a mean Full Scale IQ of 122, however, the mean education was 13.58 years. Many existing normative samples seem to include intellectually above average participants. Although education level or years of schooling are static demographic variables not affected by disease, the variability within these demographics is a potential threat. On the other hand, vocabulary or Verbal Intelligence are better indicators of BNT performance. Both vocabulary and Verbal Intelligence correlate with education, but they also vary across persons within a particular education level (Hawkins & Bender, 2002). The use of the Wechsler Test of Adult Reading (WTAR) will provide a useful instrument for this purpose for several reasons.

Given the mixed results with the current BNT research, all precautions must be taken to reduce this threat by including a normative sample that is most representative to the current US population and by using the WTAR to obtain an estimated VIQ as a basis for BNT performance expectations.

CHAPTER 13

INTRODUCTION FINAL SUMMARY

Age, gender, education, and culture have been found to be important concomitants of BNT performance. These factors appear to be correlated, and therefore, all known factors will be examined in the proposed study of naming function. Our study will utilize statistical and theoretical methods to account for possible variance in scores besides true naming ability. If age differences are found, it is suspected that differences in controlled processing are responsible for age differences in naming performances and not lexical retrieval mechanisms. If age differences are present, latencies more likely will uncover age differences in automatic processing or lexical access in addition to age differences in controlled processing following word-retrieval failure. Hence, the BNT-L will be a more inclusive measure of word finding than other instruments that measure accuracy alone.

The mixed findings in the literature appear to result from inconsistent administration and scoring procedures, unrepresentative and inconsistent use of normative samples, and many potential confounds that affect successful lexical access and word retrieval. This all presents a major difficulty in clinical practice, because the difference between normal aging and mild pathology is not differentiated which leaves a risk for misdiagnosis. The BNT's original and most useful application is for clinical populations with anomia or aphasia. The BNT-L will be constructed with the intention on assessing normal and mildly-moderately impaired individuals, providing both an index of automatic lexical retrieval functions and paralinguistic-controlled functions. By design, further studies can explore age differences in TOTs and TOT resolution following phonetic cues with the BNT-L. The fact that dysnomia is a common feature in many neurological disorders (Hawkins & Bender, 2002), the BNT-L can have wide applications and can possibly serve as a measure of recovery of functions following injury or disease (Dunn et al.,

1989). As stated previously, ardent attempts were made to create a sample of healthy adults, representative of the current US population in order to provide normative data representative of the clinical population for whom it is intended. It has been recommended that BNT normative data be moderated by estimated premorbid vocabulary (Hawkins & Bender, 2002) and other variables (e.g., gender, age, education, IQ) as extensively reviewed and discussed.

The main goal of this study is to use modern methods of item selection with data-based validation procedures for a known neuropsychological instrument, the BNT, to refine the BNT such that latency can be used as a supplementary indication of cognitive process; not only assessing an accurate/inaccurate response, but evaluating the time to process naming ability. Traditional Item Response Theory (IRT) may not be appropriate due to several qualitative features surrounding BNT item responses as discussed. Graphical and qualitative assessment of item performance is expected.

Based on the works just reviewed, and using two experimental designs the following goals will be achieved with the current study:

1) *Experiment 1*: Collect normative data that is representative of US population with expected adjustments or corrections for age or education or verbal IQ;
2) Analyze the data to determine if age-related changes are evident, and if so, analyze automatic or controlled processing involved in word retrieval, hypothesizing only the latter will be significant with later age;
3) Modify the 60-item BNT (Goodglass et al., 2001) to create a new and shortened instrument, BNT-L, that uses latencies as the dependent variable;
4) Report clinician-friendly normative data with sophisticated statistical techniques; and,

Experiment 2

5) Validate the discriminative power of the new instrument and corresponding normative data with two high-functioning groups of adults living in the community who were referred for neuropsychological evaluation: one group with “potential” brain injury and another group with “potential” dementia.

CHAPTER 14

EXPERIMENT 1 METHODS

Written proposals for the Institutional Review board (IRB) at the University of North Texas, Denton, and the University of North Texas Health Science Center were approved for this project. The present study is divided into two separate experiments: Experiment 1 consists of the creation of a new testing instrument to measure naming ability, the Boston Naming Test of Latencies (BNT-L), and Experiment 2 measures the validity of the new BNT-L.

Participants

Two hundred thirty-five healthy adults aged 18-89 years of age (average age 43) from a diverse community in north Texas volunteered for this study. Educational level ranged from grade seven to 20 years (average education 13.8). Care was taken to select a random and representative population on US demographics (see Table 1 for sample demographics). Individuals were recruited from: a large local university, a local coffee shop, fast food chains, public food shelters, downtown passerby's, and at the local senior citizen center. Only optimally healthy individuals were included and none had a history of stroke, seizures, heart attack or bypass surgery, or uncontrolled hypertension or diabetes (see Appendix A for the Health Screening Worksheet); participants were excluded from final analyses if they did not meet the health screening criteria. All participants had English as their native and primary language and none had uncorrected vision or hearing impairment. Bilingualism was noted as an extraneous variable with potential confounds when English was not the first language used (Roberts et al., 2002).

Procedures

Participants were tested individually in a quiet area by the same examiner. A structured interview was administered first (Christensen et al., 1991, plus additional items) to screen participants for exclusionary criteria. Each complete evaluation included in the sample was given a subject number and remained anonymous. Each file received an identification number between 1 and 235 to record the data; no names or identifying information was provided outside the informed consent. Information collected from each individual was:

- Demographics of age, date of birth, gender, ethnicity, and education level;
- Handedness of left, right or ambidextrous;
- Health-related issues such as: presence or absence of menopause or taking hormonal medication; presence and type of diabetes; presence of asthma; caffeine intake and amount; and, current amount of daily cigarette usage and the presence of 10-year history of smoking cigarettes (for future studies);
- Estimated verbal, performance and full-scale intelligence quotients (VIQ), (PIQ), and (FSIQ) as obtained from the Wechsler Test of Adult Reading (WTAR);
- Positive and negative health habits as measured by the Multidimensional Health Profile (MHP, 1998);
- Response time on each of the 60 items on the BNT (Goodglass et al., 2001), number of cues, predetermined qualitative codes if present; and,
- Presence or absence of a TOT state.

Instruments

Wechsler Test of Adult Reading (WTAR)

Reading recognition has been found to be relatively stable in the occurrence of cognitive declines related to aging or brain injury, and tests of reading are accurate measures of intelligence before disease (McGurn, 2004), injury, or age-related cognitive decline (WTAR, 2003). Rather than a vocabulary test, a reading recognition test is less cognitively demanding (Lezak, 2004) and has previously been shown to be predictive of BNT performance (Hawkins et al., 1993). The Wechsler Test of Adult Reading, or WTAR, provided an estimate of an individual's level of intellectual functioning. The test was co-developed and normed with the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III), enabling the WTAR to be an effective method for predicting full-scale IQ and memory performance, with prediction equations available for WAIS-III index scores. Demographic data is available specifically for use in neuropsychological cases. Our study obtained a verbal intelligence quotient (VIQ), a performance intelligence quotient (PIQ), and a full-scale intelligence quotient (FSIQ) from each participant.

The WTAR comprises a list of 50 words with atypical grapheme and phoneme translations (e.g. menagerie, liaison) and can be completed in 10 minutes. Administration involves asking the individual to read 50 words out loud from a laminated card. The examiner's record form contains correct pronunciations. The total score is the number of words read correctly. All words are administered. (See Ginsberg (2004) for a critical review of the WTAR).

Boston Naming Test (BNT)

Reliability, validity and description of the Boston Naming Test (BNT) have already been presented in Chapter 8 of the Introduction.

BNT Standard Administration

Standard administration (Goodglass et al., 2001) of the BNT to adults begins with item 30 (harmonica) and continues forward unless a mistake is made on the first eight items. If any of the next eight items are failed, reverse testing is continued from item 29 until eight consecutive responses are correct, at which point testing resumes forward again until the individual has eight consecutive incorrect responses. Credit is typically given if the person has a correct response in 20 seconds. A provided stimulus (semantic) cue is offered if the picture is clearly misperceived as something else. For example, if a subject says, "snake" for *pretzel*, the first semantic cue is **"**it is something to eat." If the item *pretzel* is still not identified after 20 seconds, a phonemic cue (e.g. "pre" for the first phonetic syllable) is offered. Credit is given only if the picture is correctly named within the 20 seconds, and thus no credit is received with a phonetic cue. The published response booklet contains seven recordings: 1) the number of spontaneous, correct responses; 2) the number of stimulus cues given; 3) the number of correct responses after a stimulus cue; 4) the number of phonemic cues; 5) the number of correct and incorrect responses after phonemic cues; 6) the number of multiple choices given; and, 7) the number of correct choices. For more detailed instructions on standard administration and scoring see the test stimulus booklet Goodglass, Kaplan, & Barressi (2001), Lezak (2004), and Spreen & Strauss (1998).

Boston Naming Test-Latency (BNT-L) Administration

Written administration procedures for the Boston Naming Test of Latency (BNT-L) used in this study are summarized in Appendix B. All responses were recorded verbatim in the response booklet that is described below and also in the procedures. Since the purpose of this study was to modify the BNT to create a new instrument using latencies to increase the clinical utility of measuring word finding in normal and mildly impaired individuals, administration

procedure of the BNT Latency test items were generally in accordance with the standard rules intended for a clinical setting. Latencies were measured in full seconds using a hand-held digital stopwatch to simulate a clinical environment.

Modifications from standard BNT administration (Goodglass et al., 2001) were several. First, administration of individual items began with item one and proceeded forward with all items presented in order to obtain as many latencies for each item as possible for each subject. Basal and ceiling levels were ignored and all 60-items were administered to each participant.

The scoring sheet was modified to contain each item listed followed by a time line written in 5-second increments with slashes indicating increments of seconds. Responses were recorded up to forty-seconds for each item. Testing began with the examiner saying "*Please tell me the most common name for these objects in a single word as fast and accurate as you can. It is important that you try to be both quick AND accurate in your responses.*" Time was measured as soon as the card was being perceived and latencies were measured by circling the number, in whole seconds, on the protocol sheet's timeline.

Because this experiment's purpose was to create a word-finding instrument with construct validity, efforts were made to determine if delayed or inaccurate responses were due to true word-finding rather than due to variables such as vocabulary or poor quality of a stimulus picture. A thorough explanation of the purpose of the study was given as part the Informed Consent procedures (i.e., to create a test to help measure word finding by the way of naming pictures) and that each participant's feedback was encouraged to aid in this process. The experiment's purpose was expounded to ensure that the stimulus pictures elicited true word-finding rather than vocabulary or something else. Each delayed or incurred response resulted in further inquiry after completing that item by asking, "Can you tell me why you responded that

way?" Nine codes were used to notate the participants' interpretation of their experience. A list of codes used during administration of Experiment 1 is shown on table 2. Some codes are analogous to those given in the standard administration of the BNT (see Goodglass et al., 2001; Goodglass et al., 1998) and other codes are comparable to codes used in previous research (for example, Au et al., 1995; Fraas, Lockwood, Neils-Strunjas, Shidler, Krikorian, & Weiler, 2002; Nicholas, Brookshire et al., 1989; Schmitter-Edgecombe et al., 2000**).**

A semantic prompt, or "stimulus cue," and "S" (Code was marked on the protocol timeline when an individual misperceived the item as representing something else (e.g. "branch" for asparagus, by offering "it is something to eat") or if it was apparent that the individual lacked recognition of the picture (e.g. "I don't know what that is"). The stimulus cues from the traditional BNT Booklet were printed in brackets under each item.

Phonemic cues were given after 30 seconds and "P" was placed on the appropriate location on the timeline. If the participant correctly named the picture following a phonemic cue, inquiry about a TOT was asked, "Was this word on the tip of your tongue?" and "Y" or "N" was coded accordingly. Phonemic cues were the underlined portion of the target word on the BNT-L Scoring Sheet as is on the standard protocol sheet. An indication of a potential TOT occurred when a participant showed verbalizations ("V" code) that accurately described the object without naming it.

A participant was assumed to have experienced a linguistic failure, or TOT, if he or she correctly responded after approximately 1.5 seconds (see Levelt, 2001 for discussion on linguistic elements) or after a phonetic cue (see James & Burke, 2000, for discussion related to TOT). The presence or absence of TOT state was asked in this experiment for reasons of ecological validity. Thus, when the participant could not describe the item, a TOT state was

considered absent, and the incorrect response was not due to word finding but rather he or she either did not know the item or did not have a semantic reference for the item. Responses coded "N" for TOT indicated items that are not be suitable in a word-finding measure for all age groups due to the semantic nature of the item.

Unlike standard administration, unlimited number of responses was permitted for each item in order to collect data for underlying reasons for incorrect responses. Incorrect responses were coded as "RN" for Related Name or "DK" for "Don't Know," and extra verbalization were noted with a "V." "RN" was coded when a person responded with a name similar to the test item (e.g. schoolhouse for house). "DK" was coded if the response was "don't know."

Positive and Negative Health Habit Questionnaire (scale from the MHP)

The third instrument administered was the Positive and Negative Health Habit Questionnaire. This questionnaire is one of five scales imparted by the Multidimensional Health Profile (MHP) and provided separate *T* scores for Positive Health Habits (PHH) and Negative Health Habits (NHH). The PHH scale comprises questions relating to exercise, regulation of weight, proper eating habits, and wearing a seat belt, etc. The NHH scale includes information concerning behavioral health risks such as poor nutritional habits, substance abuse, etc.

Statistical Methods Used for Good Psychometric Instrumentation

Canonical Correlation: Overview

Canonical Correlation (CC) is a multivariate statistic that is used when there are several continuous dependent variables as well as several continuous variables with the goal of assessing the relationship between two sets of variables. In this study, we need to determine if there is a relationship among naming latencies (called reaction times herein for statistical discussion) and naming accuracy, and a set of variables affecting naming ability (covariate set of: FSIQ, years of

education, age, and gender). CC was used to estimate the degree of relationship between the covariate set and the principal outcome variable (ranked summed reaction time). CC finds the optimal set of variable weights (multipliers) that maximizes the between set correlations (i.e., Set 1 = Ranked Summed RT; Set 2 = Covariates). CC technique is essentially constrained principal components analysis on two variance/covariance matrices. The goal is to extract "weights" or "coefficients" that create composite variables for each set, such that the composite variable correlation, for the two composites, is as large as possible. This composite set correlation is known as the "canonical" correlation, and when this correlation is expressed as a squared quantity, it estimates the shared variance between the two sets. The purpose here is to use the ability of CC to optimally combine covariates into a single value, whereupon a statistical adjustment is made (using regression residualization) on the composite variable which only contains the principal outcome variable (ranked summed RT). This technique would allow any number of background covariates to be efficiently combined in the statistical adjustment.

Canonical Correlation Formulas

The first step in this process is to calculate the composite variables for the outcome variable and the covariate set. Expressed as an equation we are estimating the w's in the following:

Wx * Xrrt = R * (W1 * Y1age) + (W2 * Y2educ) + (W3 * Y3gender) + (W4 * Y4fsiq)

this reduces to:

Xcomposite = R * Ycomposite

where: R = canonical correlation

Wx = canonical coefficient for Xrrt (ranked summed RT)

W1 = canonical coefficient for AGE

W2 = canonical coefficient for EDUC

W3 = canonical coefficient for GENDER

W4 = canonical coefficient for FSIQ

The second step involves taking the estimated W's to create the Xcomposite and Ycomposite variables. After the composites are calculated, a simple regression analysis is performed predicting the Xcomposite from the Ycomposite variable. This regression estimation will produce "slope" and "intercept" values that allow optimal linear prediction of the Xcomposite from the observed Ycomposite. However optimal this prediction equation might be, there are some potential problems for the goals in this study. Observed Ycomposite means having the individual values on the covariates (Y1age...etc), such that the CC weights can be applied to find the observed composite. In cases where this information is available, the coefficients can be applied and an adjustment made such that Xcomposite has the effect of Ycomposite linearly subtracted (i.e., regression residualization). When that information is missing (e.g., do not have FSIQ score available), adjusting on "potential" values of the unobserved covariates can still be performed, however, special care is needed in inputting what those values would have been, and without inducing bias in the observed score (here it is the Xcomposite score). In the present study, using the *rank* transformation of the scores, allows a standard metric for the N=235 cases that are analyzed, such that the mean of the variable (ranked summed RT) is 118 with a standard deviation of 67.7. Moreover, the mean and standard deviation of the ranks will be the same for both the principal outcome variable and the covariates (i.e., mean rank AGE=118; mean rank EDUC=118, etc.). This is convenient in that, in the absence of ranking information on the covariates, using the average ranking (which is in our case 118 for all of the covariates) amounts to using our "best" guess (having no other information

available to us) as to what that ranking would have been if they had been observed on the covariates. The adjustment, in this case, amounts to adjusting on the average value for the Ycomposite (covariate). Whenever covariate information is available, then the adjustments to the Xcomposite variable will be proportional to the size of the rank on each respective covariate.

As an example of how a residualization of an Xcomposite score would be calculated, we use a rank score of 120 on the Xrrt score as an example. Here, a ranking of 118 is used for all background covariates assuming that these *covariates are unobserved.* Our "best" estimates are taken from Table 11 to be shown in Chapter 15: Experiment 1 Results. In this notation, the "Slope" value is the canonical correlation:

$$Wx*Xrrt = Intercept + Slope*[\ W1 * Y1age + W2 * Y2educ + W3 * Y3gender + W4 * \ Y4fsiq]$$

Inserting coefficients give:

$$(.000006) * (120) = .0008 + \ [.40 * (.0000115 * 118 \ + .00000043 * 118 + .00000024 * 118 + .00000560 * 118]$$

This reduces to observed Xcomposite predicted by best estimate from Ycomposite:

Observed Xcomposite		Best Estimate from Ycomposite
.00072	=	0.001638744

Residualized score is calculated then by subtracting observed and predicted:

Xresid = -0.000918744

Calculation of Standardized T-scores from Residualized Scores

Residualized scores do not necessarily have a variance of one or even a mean that is exactly zero. Our goal is to place the residualized scores on a metric that has been consensually agreed upon by previous researchers. Some examples of these scale changes are the so called

"standard Z-scores" or "T-scores." Equating scores by a change in location and scale (mean and standard deviation) involves first converting to a standard score by subtracting the scores by their mean, and then dividing the scores by their standard deviation. This results in a set of scores whose mean is zero, and whose standard deviation is one. It is important to note here that this change in location and scale does not change the underlying probabilities attached to the transformed scores – Z-score do not change the shape of the underlying probability distribution. Next, changing standard scores to a different location and scale involves adding a mean and then multiplying by the standard deviation or scale parameter. For example, a hypothetical score and change to a T-score will be calculated as:

$$\frac{(X - Xmean)}{Xstdev} = \text{Z-score} \quad \text{where mean} = 0; \text{ standard deviation} = 1$$

Next, to convert to T-score (for example):

(Z-score + 5) * 10 = T-score where mean = 50; standard deviation = 10.

The properties of these residualized scores converted to T-scores should be that they are not correlated at all with the background covariates, however, they should still be substantially correlated with the original outcome scores (i.e., ranked summed RT – Xrrt; as will be seen on Table 7 in Chapter 15: Experiment 1 Results).

Bootstrap Resampling Scheme

The name "bootstrap" refers to the analogy of pulling oneself up out of the mud by one's own bootstraps. Bootstrap scheme is a computer-intensive "resampling" method for estimating the variability of statistical quantities and for setting confidence regions that was introduced into statistics by B. Efron in 1979. Bootstrap resampling is used often as an alternative to inference based on parametric assumptions when those assumptions are in doubt, such as the case in the current study of not knowing the underlying distribution given the multiple effects on naming

reaction time. The concept of bootstrap is that, in the absence of other information, the sample itself offers the best guide in sampling distribution. In the present study, 235 records (N=235) were sampled including the variables of reaction time, age, gender, education, and FSIQ from the original data set using "replacement." In essence, this amounts to sampling from the integers 1 to 235 with replacement (and equal probability of selection) and using these numbers and indices to pull records out of the original data set to form a new data set – which is then called a "bootstrap sample." Then, the empirical distribution of our estimator in a large number of bootstrapped samples is used to construct confidence intervals and tests for significance. It is important to distinguish between randomized and bootstrap samples. Randomized samples are generated by scrambling the presented data (sampling *without* replacement) whereas bootstrap samples are generated by sampling *with* replacement from the original sample. Therefore, some data points from the original sample are expected to be present two or more times while others are absent. For example, resampling the integers 1 to 235 with replacement produces the following potential record indices: 99, 61, …..174, 151, …..61, 135, …..etc. for a total of N=235 indices, with 61 noted as twice sampled because we are sampling with replacement. These index numbers refer to the case number in the original dataset. Once finished resampling, a new data set with slightly different sample characteristics reflect the sampling process. In effect, the sample is now treated as a population and the resampled data is treated as new sample; the relationship between the original sample and the true population is reflected in the relationship between the original population and the resampled data. This property has been referred to as the "plug-in" principle by Efron and Tibshirani (1993). In a typical bootstrap resampling process, many bootstrap samples of some specific sample size are created (for this study the sample size is 235). The literature indicates that for the number of variables that we have those 1000 bootstrap samples of

N=235 should be sufficient to estimate the bias in the parameters of our canonical correlation model. Less biased estimates of model coefficients are produced by estimating the model many times with bootstrap samples, and then averaging the bootstrapped model parameter estimates to get a single less biased bootstrap estimated model coefficient. Standard errors of these bootstrap coefficients are obtained by taking the standard deviation of all of the bootstrap parameter estimates. The standard error is essentially the variation in the distribution of bootstrap estimated coefficients. Such that, 2.5th and 97.5th percentiles can be calculated from the distribution of bootstrap estimated coefficients to produce nonparametric confidence intervals on the bootstrap averaged parameter. Less biased versions of this simple percentile bootstrap exist, but are more complicated for the purposes of this study.

In the present study we approximate the external validity of the modeling process and the subsequent estimated coefficients by resampling from the original data and applying the bootstrap methodology. For each bootstrap sample: ranks are assigned, a canonical correlation is performed, and a subsequent regression is performed to get the "slope" and intercept" that predicts the Xcomposite from the Ycomposite. All bootstrapped coefficients are averaged to provide a bootstrap mean value for the coefficient and a bootstrap standard error for the coefficient. To provide further generalization on this process, the overall described bootstrap process was replicated itself 100 times, so that in essence, each bootstrap mean is actually based on 100,000 bootstrap samples. This second "loop" or "tier" provides another set of averages and errors for the first set of bootstrap average coefficients and standard errors.

Development of a Measurement Model

Latent variable modeling and item analysis require that certain assumptions be made concerning the population and the sample. The more important assumptions for linear factor

analysis can be minimally broken down into at least three interrelated assumptions: 1) Errors (or residuals) are normally distributed in the population; 2) The relationship between the latent (or unseen) trait scores and the observed scores are functionally related in a specific way – linear (e.g., straight line relationship with two parameters: slope and intercept); and, 3) That observed score variance is additively summed as measurement error and true score variance, where the true score variance is a result of individual's differences on the measured trait. These three major assumptions create a practical priority early in developing a statistical measurement model of the data.

An immediate concern for factor analytic methods is the selection of an appropriate measure of inter-item association. Pearson correlations will give the highest estimates of association (also less biased as estimates of population association values) whenever a bivariate normal distribution best describes the errors or residuals around a linear function that relates the two variables, say X and Y. If two variables are monotonically related to one another (as in an exponential curve or logarithmic curve), and contain order (or rank information) that is of primary interest, then the Spearman Rank correlation is an appropriate measure of association for the pairs of items. Quite frequently, Pearson correlation measures and Spearman correlation measures will be almost identical whenever normality assumptions are met on the residuals (in largish sample sizes ~ N>100). Where this is not true is when there are outliers present in the outcome or predictor variable; Pearson correlation and standard linear regression are known to be highly non-robust to outliers. Additionally, measurement error affects Pearson correlation and linear regression estimates by reducing the estimates of the sample correlations and beta coefficients.

Other sources of error and bias include censoring or truncating the observed variable by some unseen mechanism. This can lead to missing values in the worst case, or can lead to limited or truncated values due to some coarse measuring technology (e.g., using second-level accuracy from an analog stop watch versus using millisecond accuracy from a digital stopwatch). "Ceiling" and "Floor" effects are a kind of truncation on data where the natural variation of a variate is constrained such that potential values on the variate are either practically impossible or are logically impossible to record. It is known that rank transformations, where continuous data is concerned, can be very helpful in alleviating many problematic assumption violations. For example, rank transformations are known to reduce the effect of outliers on many parametric statistical techniques, in essence creating robust versions of the parametric technique. Rank transformations are also used in ANCOVA designs to reduce the impact of measurement error on covariates that are used for score adjusting or group equating on background covariates. Also, the use of rank transformations in ANCOVA allows statistical adjustments to be made even when the basic "homogeneity of regression" assumption has been violated.

Other measures of association that are relevant for measurement models are the "threshold" parameter variations of the Pearson correlation measure for dichotomous or polytomous ordinal data – the so called tetrachoric correlation or more generally the polychoric correlation measure. These correlations assume an underlying latent continuous population for the variates, but that some censoring occurs due to some unseen threshold for an item. These thresholds are estimated as parameters, and using these thresholds, Pearson correlations are calculated that have been adjusted for the censoring – in effect this corrects the bias that occurs as a result of the censoring.

Confirmatory Factor Analysis

Once an appropriate measure of inter-item correlation have been selected, then appropriate factor analytic strategies can be used to extract modes of covariation from the variance/covariance matrix of all of the items under study. These modes of covariation are called "factors" (assuming a measurement error model) and represent the common trait as measured by the set of items constructed. Item error or the "uniqueness" of the items contribute to non-covarying variance in the variance/covariance matrix (or correlation matrix) and result in smaller amounts of variance in the trait as measured in common by the items. These item uniquenesses (or residuals) lead to less precision in our estimates of how well the items measure true variance, or the trait being measured. These estimates of how well the items measure true variance are called "factor loadings." Large factor loadings on an item indicate that that an item measures, or accounts for, a large percentage of true variance in the observed score variance for that item. If that item also has a low uniqueness, or low item error, then this item is contributing greatly to the overall reliability of the summed score across all items.

Particularly important to this present study is the use of "confirmatory factor analysis" (CFA). CFA allows hypothesis testing of a particular factor structure(s) with regard to which items comprise that factor. For example, it is of interest to know whether the 15 selected from the BNT for a short form of the BNL-T, can still be considered to be homogenously measuring the same "domain" content that they were constructed to measure (e.g.. semantic labeling of everyday objects). Or in other words, this study aims to use CFA to not only estimate the true variance and measurement error of the selected items (e.g. reliability), but also to "confirm" (or test the hypothesis) that the remaining 15 selected items can be reasonably said to comprise a single set of items homogenously measuring a single trait or single factor. CFA can also be used

to provide information on the reliability and validity of a set of items. Factor loadings estimate the true score variance in a classical measurement model and the uniquenesses estimate the measurement error. Reliability indices can be estimated directly from these factor loadings and uniquenesses (McDonald, 1999). Furthermore, some reliability estimates, based on "congeneric measurement" factor models, can be considered "validity" coefficients – a correlation between the sample items and infinite domain of all possible items that could have been used (sampled). The omega coefficient is one such reliability coefficient (McDonald, 1999). Cronbach's coefficient alpha, a common reliability estimate, makes assumptions that the inter-item correlations are reasonably homogenous, whereas the Omega coefficient does not make such an assumption; in omega coefficient error variances and inter-item covariances are not constrained. As such, the Omega coefficient is lower bound on the true population reliability. Practically, this means that the Omega coefficient will always be equal to or larger than coefficient alpha.

Connections between Factor Analysis and Item Response Functions

For purposes of the present study, estimated loadings and uniquenesses allow for statistical evaluation of an item's performance in the overall summed score performance, for example, the summed reaction time score on the 15 items selected. In fact, a correlation between the summed score and the performance on a single item is a rough estimate of the factor loading for an item (for loadings in standardized form – correlations). The loading can be considered as the extent to which the observed score is predicted by the underlying latent trait or factor, essentially a slope coefficient. The larger the slope coefficient, the greater the range in the observed score performance as predicted by the underlying differences in the trait amongst individuals (common latent trait or factor). In item response theory these parameters (loadings) are called "discrimination" coefficients because item response functions with higher loadings, or

discrimination coefficients discriminate well between individuals of high and low ability (e.g., allow error-free predictions in observed performance as predicted from an individual's estimated latent ability).

CHAPTER 15

EXPERIMENT 1 RESULTS

The sample's demographic information of age and gender was somewhat comparable to that of the 2000 United States Census data (US Census Bureau, 2000; see Table 1), however, education and FSIQ were slightly higher in this sample. Item variability as measured by histograms showed individual variability but not much item variability indicating that the items as measured with latency had discriminative potential and that the data was conducive for further item analyses.

Statistical analysis used R as the language and environment for statistical computing. R is a GNU (Acronym for "GNU's Not Unix") project, which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

Overall Accuracy

The mean for the overall accuracy of the reaction time data (with this study's administration procedures) was slightly higher than that of published BNT norms (M = 57.85). Because latency time continued when standard administration procedures would have ceased (i.e., would be deemed "inaccurate" without another chance for a correct response), data was also analyzed removing items that would have been an incorrect response had standard scoring applied. The overall sum of correct scores using standard administration and scoring procedures were very comparable to published BNT norms, all adult ages (M = 54.01our sample, BNT norms, M = 54.00, Goodglass et al., 2001).

Item Analysis

Codes

All stimulus items that had a code (input numerically 1-9 corresponding to how they are listed in Table 2; for example, RN=1, SS=2, BP=3, etc.) were extracted and the numerical code was converted to string variables (to permit separated characters; for instance, 1:3 string shows a code for RN and BP for that item) before analyses were run.

Creating a Shortened Version

Selecting the most discriminating items for the final version used a descriptive ordinal approach to an Item Response Theory (IRT) curve for reaction time data. Discriminating items were items in which theoretically, one-half the sample did well and one-half did not do well. Items showing discrimination across groups began by first "blocking" the sample into 16 blocks (with 14 in 15 blocks, and 15 in one block) followed by a "smoothing technique" to reduce variability and indicate clusters, or create IRT curves, to analyze the observed data. The data was "smoothed" by taking the average class reaction time by accuracy in each class and created into continuous data. Data was analyzed numerically and pictorially through computer-generated item response curves ranked ordered for each item (those centered in the middle with a steep slope showed half of the sample performed well and about half not well). The blocked sums of the means for each item's reaction times were then rank ordered and correlated using Spearman's method. The sequential steps for selecting final items were as follows:

1) Select items with accuracy by reaction time correlation between $r = .50$ and .75 to exclude items with low variability within responses (i.e., to select discriminating items where not everyone did poorly and not everyone did well).
2) Eliminate items that had frequency of RN code > 20%.

3) Eliminate items that had frequency of SS code > 20%.
4) Eliminate items that had frequency of BP code > 20%.
5) Eliminate items that had frequency of SW code > 20%.
6) Eliminate items that had frequency of DKW code > 20%.
7) Eliminate items that had frequency of VW code > 20%.
8) Eliminate items that FSIQ r > .33.
9) Consider all correlations of stimulus picture and gender, education and age.

Stimulus items were then selected if their accuracy by reaction time correlation was between .50 and .75 to maximize discrimination ability (i.e., selected items with high variability in responses). This resulted in excluding 37 of the 60 items due to poor variability in individual responses for these items. Next, all 60 stimulus items were analyzed according to the codes appointed during administration from participant feedback (refer to Table 2 for a code descriptions) to evaluate subjective accounts of what some stimulus items were assessing as well as frequency of paraphasias and presentation of stimulus or phonemic cues or TOTs for that item**.** Table 3 shows the 60 items and the frequency of each corresponding code.

Several stimulus items generated no comment from the participants and had no code. Stimulus items that individuals did not know are indicated with code 7, DKW. If the person felt that vocabulary was being assessed (code 8 = Vocabulary Word) indicate items that were assessing other cognitions than naming abilities. Items deemed to generate a poor visual percept were coded as 3 for Bad Picture (e.g., e.g., 45 individuals considered "tripod" a bad picture, most perceiving a compass; and, 15 people initially saw a branch instead of "asparagus"). Several items had multiple codes showing several participants had similar comments for that item (e.g., 17 individuals produced a phonemic paraphrasia such as terrace for "trellis" and 71 individuals

did not know this word entirely). Common strings of codes were in the data that warranted further review so a table was made to indicate common strings for individual items. Tables 4a and 4b show frequent code strings that occurred and the stimulus item with the frequency for that item.

The most frequent code sequence was 6-9 (Phonemic Cue and Verbalization) with 20 items coded with this string. Another code sequence with high frequency of 18 occurrences was 5-6-7-8 (stimulus cue – phonemic cue – don't know word – vocabulary word). Individuals used this code sequence when the stimulus item was not part of their knowledge base (e.g., 53 individuals in this sample reported this code sequence for "trellis"). A similar code string includes the string above with the addition of 9 (verbalization) which could implicate the stimulus item was sufficiently part of their knowledge base to verbalize around the target word but the specific word was not part of their working vocabulary.

Response styles were not significantly different between younger and older individuals, showing comparable codes through all age groups. Additionally, 98% of tip of the tongue phenomena was resolved following a phonemic cue, with no differences between younger and older participants.

Selected 15 Items

The following 15 items were selected as having the most discriminative potential as being most likely to measure naming ability after careful scrutiny using the qualitative coding and nine step exclusionary criteria listed above: Helicopter, Hanger, Racket, Snail, Seahorse, Dart, Rhinoceros, Acorn, Igloo, Cactus, Escalator, Harp, Pyramid, Funnel, and Asparagus.

Examination of Distribution Characteristics on 15-item BNT-L

In general, data distributions in this study were clearly non-normal. This non-normality ranged from being highly skewed, in the case of summed reaction time (RT), to being mildly non-normal, in the case of education levels (EDUC) (See Figure 2.). In the case of bivariate relationships, non-normality on the univariate distributions do not necessarily lead to curvilinear (nonlinear) relationships in the scatter plots, but almost always do if the skewness on the univariate distributions are directed in opposite directions on each respective variable's scale. For example, if the bulk of the data for X is concentrated on the low end of the x scale, but the bulk of the data for Y is concentrated on the high end of the y scale, some form of curvilinearity will likely be detected, if only in graphical form. Likewise, the data are likely to appear "heteroscedastic" if univariate data distributions are skewed in the same direction on their respective scales. In this case, data concentrated in the common heavy region will "tail" off in the direction of the common region of the "light" tails of the univariate data distributions (a good visual description would be a tear-drop shaped data clump in the scatterplot).

Figure 3 is a plot of the summed RT scores plotted against the summed accuracy scores on 15 items that were eventually selected for the short form of the BNL-T (including all 60 items in bivariate plots of items produced essentially the same results). In Figure 3, the dotted straight line presents the line of best fit for a linear function. The curved line is a scatterplot smoothing curve. The scatterplot smoothing curve is a device that aides researchers in detecting most forms of nonlinearity in the scatterplot. Whenever the dotted line and the solid (curvilinear) line overlap, there is reason to believe that a linear relationship will be sufficient in describing the functional relationships for the two variables. It is clear from the Figure 3 scatterplot that there is substantial curvature in the bivariate relationship and opposing skewness in the univariate

distributions. Accuracy is nearly maximized and pushed up against a ceiling effect of 15 on the high end on the X axis, and RT is mostly scored at 1 or 2 seconds at the low end of the Y axis. These are variations that cause the Pearson correlation to be substantially biased downwardly, or in some cases to have "sign" reversals in the correlation where none would be expected. Figure 4 displays the same relationship between RT and Accuracy after rank score assignments have been made separately for each of the two variables, summed RT and summed Accuracy. The dotted line and the solid line practically overlap. This would suggest that a Spearman correlation (or a corresponding Pearson correlation on the ranks) would maximize the measure of association between accuracy and latency. Pearson correlations and Spearman correlations were compared on all pairs of variables, and unwaveringly the Spearman correlations were larger than the Pearson correlations. Therefore, for most results in the present study, Spearman correlations can be assumed accept where indicated otherwise.

Table 5 displays the Spearman correlations of the summed RT scores with person level background covariates. These correlations indicated correlation of ranked summed RT with AGE (r = .15); EDUCATION (r = -.2197); GENDER (r = .1307); and FSIQ (r = -.41).

Reliability of Summed RT Scores and Ranked Summed RT Scores

The reliability of the Summed RT Scores was assessed by performing a Maximum Likelihood Confirmatory Factor Analysis (CFA) on the 15 selected items. Two separate CFA analyses were performed: 1) CFA for on the summed RTs (non-ranked) for the selected items; and 2) CFA on the summed and ranked RTs for selected items. In general, the CFA indicates that the loadings are higher; uniquenesses smaller; and reliabilities higher for the ranked scores. The results from the CFA on ranked scores are presented in table 6. Reliabilities were assessed on both the ranked and unranked scores. The Omega coefficient (McDonald, 1999) is contrasted

against Cronbach's Alpha coefficient for both the ranked and unranked scores. In general, the Omega reliability coefficient is considered to be a better estimate of the underlying population reliability whenever pairs of items cannot be considered to have equal variances or covariances (i.e., lower bound on the true population reliability). Whenever the variance/covariance assumptions are met, the Omega coefficient will be equal to the Alpha coefficient. The Omega reliability coefficient is .71 for the ranked RT sum scores.

Loadings ranged from the smallest value of .172 for item FUNNEL, with the largest loading for item IGLOO - .73. Loadings were higher for the CFA on the ranked RT sum scores. These higher loadings, taken with the higher reliability estimates, provide further evidence that ranking the RT sum scores improved the measurement properties of these selected items.

Model Fit: Maximum Likelihood Confirmatory Factor Analysis

Overall model fit for CFA are indicated by a number of different criterion taken together collectively. Overall model fit for CFA is assessed by using a "goodness of fit" index (GFI). Rules of thumb for these types of indices are to look for GFI that are greater than or equal to .90. A value of .90 indicates a "excellent" fit of the assumed model to the observed data. Both the GFI and the adjusted GFI (adjusted for sample size) indicate that a 1-factor model for the 15 items is an excellent fit. The RMSEA (root mean squared error) index is considered one of the more important indices for assessing model fit. The rule of thumb for the RMSEA is that the RMSEA should be less than or equal to .05. The .04 value of RMSEA indicates an excellent fit in the 1-factor model of the 15 items. The 90% confidence intervals for the RMSEA suggest that there is high precision in this estimate of .04. The suggestion is that in 95% samples with similarly constructed confidence interval (CI), an interval of [.026, .060] will contain the true

population value of the RMSEA for the fitted model. While an estimate of RMSEA values < .05 are considered excellent, RMSEA values that are high as .08 can be considered adequate.

For large sample sizes, the Chi-square value for the model fit will almost always be small (in this case Chi-square = 130.42, p=.003). This should NOT be taken as an indication of poor model fit, since unadjusted Chi-square indices are highly sensitive to large sample sizes, tending to suggest model rejection with small differences between data and model.

Lastly, the Baysian Information Criterion (BIC) is a relative measure of model fit used to compare either nested or non-nested competing models. BIC values that are small are to be preferred over larger BIC values. In comparison, the BIC for the Spearman-based CFA is greatly smaller compared to the Pearson correlation-based CFA: -360.94 for the Spearman CFA and –71.016 for the Pearson CFA (note that smaller means a larger negative value here). In general, the Spearman-based CFA is superior to the Pearson-based CFA (only partial results of the Pearson based CFA are included for comparison).

CFA Loadings on 15 Selected Items

For the Spearman-based CFA, loadings ranged from the smallish value of .17 for item FUNNEL to a largish value of .730 for item IGLOO. Figure 5 displays the item order plotted against the size of the loading. It is clear from this figure that as items increase in difficulty (see Figure 6 plot of accuracy by item number), that CFA loadings initially increase, then decrease again. A better representation of this effect is in Figure 7. Figure 7 displays the "information" contributed by each item to the overall reliability of the ranked summed RT score. Information is the ratio the squared loading divided by its error variance (uniqueness).

The same pattern of an increase then a decrease is also evident in the sequence plot of the information for the items. The pattern can be understood when looked at from the perspective

that items that are too easy or are too hard provide very little discrimination across ability levels. Item labels that are in the range of 25-45 are in the range of medium difficulty on accuracy (Figure 6), and are also the items that provide the greatest test level information (e.g. ranked summed RT score) contribution. This is also understandable from the standpoint that RT and accuracy are highly correlated (see Table 5). As such, accuracy and RT are both measures of "cognitive load" and should display similar item response characteristics. As items become more difficult (assuming participants are following instructions properly), then RT values should increase. Correspondingly, as items become easier, RT values should be smaller. Another view of this is displayed in Figure 8.

Figure 8 is a graphical representation of trivariate relationship between three quantities: ability (as measured by accuracy - X axis); latency (as measured by RT – Y axis); and, item type (as measured by the coloring of the item response lines). As discussed in the section above Item Analysis: Creating a shortened version, overall summed accuracy on all 60 items was blocked or binned into 16 ability classes (or ability levels). To create these ability classes, participants summed accuracy scores were sorted and then classified into 16 ability levels ranging from low to high (0 percent to 100 percent accurate). For each ability class, the average RT value in seconds was plotted on the Y axis for all 60 items. These RT response functions were then color coded in sequence from 1 to 60. The sequence of colorings was displayed in the legend along with the item number (upper right hand part the panel).The RT response functions were the data based equivalent (as opposed to model based) to item response functions (or item response curves) in Item Response Theory (IRT).

In IRT modeling, we look for items which have steep slopes (large loadings or discrimination coefficients). These steep functions imply that for a difference on the ability

continuum (X-axis), that a corresponding change can be detected in the continuum that depicted on the Y-axis – here is the average latency). The difference here, in comparison to IRT curves, is that this RT Y-axis depicts an observed response, not a latent response. IRT curves depict the latent probability of a correct response on the Y-axis. To create such a curve that depicts the latent ability on the Y-axis, we would need to plot the average "factor score" for each ability class on the Y-axis. While the CFA analysis that was performed does provide these estimated factor scores for each participant, this study is primarily interested in item selection, whereby the selected items still allow for high correlation between the ability estimates (accuracy) and the latency estimates (RT) values. It is assumed that accuracy is an observed estimate of latent ability estimates. The goal is to select items that still allow for discrimination capability between accuracy and RT (i.e., predicting RT from accuracy and vice-versa), and also allow for discrimination between observed latency, and whatever latent trait that latency measures on these items (i.e., high loadings on the Spearman-based CFA). It is clear from these results that each of the items exhibit reasonable measurement characteristics; the selected items provide good "information" toward the summed score and valid content for the domain being measured. Take together as a whole unit, and especially after considering the qualitative analysis that contributed to the selected items, these 15 items demonstrate acceptable reliability and have a valid factor structure for the domain under consideration (semantic naming).

CHAPTER 16

EXPERIMENT 1 DISCUSSION

The BNT, although designed as an aphasia screening test, is widely used as part of more general batteries used to diagnose more general forms of cognitive decline. While there is absolutely no reason to expect that measurement of response latencies as part of the ordinary administration of the BNT (i.e., using whole second timescales) would increase the test's sensitivity to word finding problems—where the process of lexical access, from object recognition through phonological synthesis either succeeds or fails within the first two to four hundred milliseconds of stimulus presentation—our hypotheses was that measuring the average length of time required for an individual to recover from such failures might provide meaningful discriminators for earlier detection of general cognitive decline.

A review of previously published literature surrounding naming ability, factors affecting naming ability, and the Boston Naming Test revealed complex information to consider when modifying the BNT to measure latencies. In *Experiment 1*, 235 optimally healthy individuals, randomly selected and fairly representative of the current US Census, completed the most recent 60-item version of the BNT (Goodglass et al., 2001) measuring latencies of response times. From these, fifteen items that were considered to show good individual variability and discriminative power were selected to comprise a new instrument, Boston Naming Test with Latencies (BNT-L). Qualitative features of the items were coded in the data collection process that served to greatly facilitate selection of fifteen items that were most pure to the assessment of naming ability rather than measuring other cognitive functions such as vocabulary or visual perception.

It was hypothesized that latencies would permit a more sensitive measure of age-related difficulties than other studies that not show changes, or changes were not evident until an

individual was in their 70s or 80s. Our overall results of *Experiment 1* showed no difference between older and younger individual's speed of responses on the BNT, and even slightly better response time as an individual age. Both of these findings were consistent with previous work that indicated that naming ability was similar in young and older adults (Cruice et al., 2000; MacKay et al., 2005) and that naming performances improved with aging (Farmer, 1990; Schmitter-Edgecombe et al., 2000; Thomson & Heaton, 1989). Analyzing different response styles through the use of pre-established codes between younger and older age groups also proved to be insignificant in our sample, showing comparable response styles across all ages. Older people's response styles, such as verbalizations or picture descriptions, were also similar to younger people which are contrary previous literature (Nichols et al., 1985; Obler & Albert, 1985) that found differences in response style, particularly with older individuals showing more circumlocutions and picture descriptions.

Lack of differences found between younger and older participants' BNT response latencies is likely partly due to careful consideration with creating our normative sample. Only optimally healthy individuals and those with more positive health habits compared to negative health habits were included. Health habits were considered in conjunction with a health screening worksheet that indicated confirmed disease states. An individual's Health habits were measured for several reasons. One, an individual's health has been shown to affect naming ability (Hickman et al., 2000), and secondly, not everyone has knowledge or confirmation of having diabetes or high blood pressure, and thirdly, this information will be used for future studies assessing naming ability and health habits.

Selecting only healthy individuals is helpful when trying to assess natural age-related declines in naming ability, but an optimally healthy group may not be truly representative of the

clinical population where it may be used. Careful consideration of the compatibility of the clinical population with the sample from which the normative data was obtained is critical to reduce Type I and Type II errors (see Ross & Lichtenberg, 1997 for examples). Our population sample was obtained in a college community where individuals with higher intellectual interests and abilities may more likely reside than other locations. Despite great effort to minimize this effect by collecting data from diverse samples, the normative population group comprised individuals with somewhat, though not significant, higher education levels and full scale IQ's than comprising the most current US census.

The number of words named following phonemic cueing has been a useful indicator of verbal retrieval, and its influence on everyday conversations (Lezak, 2004). The response pattern of response accuracy following a phonemic cue was evaluated and confirmed in the current study. When controlled processes were assumed to be used, or when latencies exceeded two seconds, participants were asked directly if they were experiencing a TOT. Resolution of a reported TOT occurred within seconds of receiving a phonemic cue as hypothesized and reported in the literature (Lambon Ralphe et al., 2002). The normative group averaged similar increases in response times following a phonemic cue similar to increases reported by Lansing et al. (1999). No difference was found between older and younger adults in their ability to resolve TOT's; older and younger adults were equal in their ability to resolve TOTs which supports the literature specific to this area (Burke & Nix (1996).

This study used the most current edition of the BNT (Goodglass et al., 2001). Despite the majority of research in the literature has used earlier versions, changes did not appear significant, nor did our findings indicate otherwise. Usage for the current edition should increase as more empirical studies validate the new edition (Lopez et al., 2004).

Given the large amount of statistical difficulties with analyzing this type of data, it appears that measuring latencies in whole seconds may not be sufficient to assess naming ability. Measuring whole seconds was thought to be sufficient to assess lexical failure and would most simulate a clinical environment. However, lexical retrieval is a very rapid process and using the nearest whole second possibly loses much valuable data. The newer stop watches are capable of measuring in 10^{th} of seconds and future research on BNT latencies are encouraged to use smaller increments of time to increase sensitivity of this method to assess naming abilities. More precise measurement of latencies could possibly uncover reasons for the equivocal findings with subjective and objective reports with diminishing naming abilities with advancing age.

A major criticism with on psychometric studies of the speed-accuracy relationship is the lack of distinction between latencies collected on items correctly and incorrectly solved. Mean latency time are not always valid (Vigneau, Blanchet, Loranger, & Pepin, 2002), however, using ranks addresses this psychometric issue because an individual item's result is based on a rank rather than part of a mean score that, if incorrect, would greatly affect overall score. Regression analysis was necessary to consider the confounding variables of age, gender, education and IQ as reviewed in the literature and as relevant in our data sample.

This experiment was followed up by *Experiment 2* to assess the validity of using the BNT with latencies in a clinical environment. It is hypothesized that measuring BNT latencies will be more sensitive to detecting naming difficulties than accuracy alone in a population where impairment is not obvious. Two separate referral reasons will be used to determine if type of impairment makes a difference in using the BNT-L when assessing highly functioning community-dwelling adults.

CHAPTER 17

EXPERIMENT 2 METHODS

Participants

Validation Groups

Participants in the validation group were individuals who were living independently in the community and were referred for neuropsychological assessment for one of two purposes. The BNT-L was administered as part of an established battery of neuropsychological tests. Individuals in Validation Group 1 were referred for neuropsychological for purposes of litigation to assess possible neurocognitive impairment subsequent to an automobile accident that occurred one or more years previously. Individuals in Validation Group 2 were referred to a health clinic for geriatrics to assess the presence of cognitive impairment, with suspected Mild Cognitive Impairment, dementia of Alzheimer's type or dementia with vascular etiology. All ethnic groups were included and participation was voluntary for both groups. All participants had English as their primary language and all had correctable vision and hearing. Exclusion criteria for this group of individuals were: obvious presence of aphasia, impaired visual perceptual abilities, and lack of effort.

Procedures

Instruments

Each participant in the Validation Group 1 completed four measures: The Boston Naming Test of Latencies (BNT-L); the Wechsler Test of Adult Reading (WTAR); Hooper Visual Organization Test (HVOT), and the Test of Memory Malingering Test (TOMM). The two additional tests were included for this sample (HVOT and TOMM) that were not in the testing for the normative sample, and the Multidimensional Health Profile (MHP) was not administered to the validation group as the purpose of the MHP data was to assist with creating the normative

data and for future research. Validation Group 2 was administered the HVOT but the TOMM was not administered as this was not part of the clinic's assessment battery, and there were no questions regarding reduced effort in this sample.

BNT-L

The BNT-L has been reviewed above in *Experiment 1* and administration and scoring instructions are presented in Appendix B. The same administration and scoring rules applied to both populations in the validation group as those in the normed population.

WTAR

The WTAR has been reviewed above in *Experiment 1*. Administration and scoring rules were the same for the validation groups and the normal populations.

Hooper Visual Organization Test (HVOT)

The Hooper Visual Organization Test (HVOT) was used to screen for visual analytic abilities, which could affect an individual's ability to perceive the BNT-L stimuli. Inclusion of participants with this impairment type could inflate latency times and improperly suggest naming impairments which may be only secondary to visual perceptual problems of test stimuli. The HVOT was designed as a brief screening instrument to measure the ability to conceptually organize visual stimuli (Hooper, 1983). The test contains 30 line drawings of simple objects that have been cut into pieces and arranged in a puzzle-like fashion. The task requires correctly naming an object when pieces are arranged correctly; each item receives a score of 0, .5 or 1. Performance depended upon "visual analytic and synthetic abilities, and on the capacity to label objects either verbally or in writing" (Hooper, 1983, p. 1.).

A Total HVOT Score was obtained after adding the number of correct responses, with a maximum score possible of 30. More than 11 incorrect responses are indicative of organic

pathology (Lezak, 1995) and individuals in our validation sample who missed more than 10 were excluded from the study based on poor visual analytic abilities.

Test of Memory Malingering Test (TOMM)

The Test of Memory Malingering (TOMM) was designed to help distinguish individuals with true memory impairment and malingers (Tombaugh, 1996). For our purposes, the TOMM was used as a measure of effort, to determine if the individual being tested was putting forth his or her best effort in their responses. This is necessary in the forensic setting. The TOMM consists of showing participants two sets of 50 pictures followed by recognition trials after each set. Data is reported for cognitively intact adults, those with cognitive impairment, aphasia and traumatic brain injury. Individuals in our sample who scored less than 45/50 (90%) in the second trial were excluded from further analyses due to the potential of poor effort that could invalidate their responses on other measures given in the study. This cutoff is suggested to identify normal (Lezak, 2004), litigating, and nonlitigating (Rees, Tombaugh, & Boulay, 2001) individuals using suboptimal effort.

Because the TOMM stimuli are similar to that of the BNT, Tombaugh (1996) recommended that the BNT be administered first to avoid contamination of TOMM memory items. The TOMM is relatively unaffected by age, education, or moderate cognitive impairment (Lezak, 2004).

CHAPTER 18

EXPERIMENT 2 RESULTS

Calculation of Residualized T-scores for the Normal Comparison Group

As described in the methods section, canonical correlation and simple regression were used to create a transformed score for the primary variable of interest, the ranked summed RT. The first step in this process was to use canonical correlation to calculate the optimal weights that relate the ranked summed RT to the background covariates: AGE, EDUC, GENDER and FSIQ. Another goal in Experiment 1 was to apply modern methods that produce less biased coefficient estimates. A Bootstrap resampling algorithm was used to adjust the coefficients downwardly so that the adjustment procedure (residualization of the ranked summed RT score) produced less biased adjusted scores. Table 7 presents the results of estimating (and calibrating) the coefficients of the following equation:

Wx * Xrrt = Intercept + Slope* (W1 * Y1age + W2 * Y2educ + W3 * Y3gender + W4 * Y4fsiq)

The column labeled "Bootstrap Mean of Coefficients" are the result of averaging 10,000 canonical correlation estimates of resampled data sets (with replacement) from the original set of N=235 cases. These coefficients were placed in the equation above and the resulting equation was used to residualize the observed Xrrt score – the ranked summed RT score. These residuals had a mean = .00001441024, and a standard deviation = 0.0003757351. Using this mean and standard deviation, the residualized scores were converted to Z-scores first, and then the Z-scores were converted to T-scores. The resulting percentiles for the adjusted scores, scaled as T-scores, for the normal comparison group, are given in Table 8 (will be the final, step 3 in BNT-L scoring procedures).

Table 8 contains the final set of comparison scores to which any incoming ranked summed RT score is to be compared (after adjustment). Since any new, incoming summed RT

score (e.g., test case) will get a ranking assigned to it in comparison to the normal group, it is possible to take all possible rankings (1-235) and create a set of predetermined T-scores for comparison, based on the adjusted rank. Table 8 displays these predetermined RANKS to ADJUSTED T-SCORE assignments. Before Table 8 can be used, an "incoming" summed RT score must have a rank assigned to it. This rank must be based on the percentile standing of the incoming score to the original (un-ranked) summed RT score data of all 235 scores of the normal comparison group. This rank can then be calculated by multiplying 235 times the percentile standing found in Table 9.

The 4-step scoring BNT-L procedures can be better explained with an example. An individual scores 48 on the BNT-L, summing responses from all 15 items totals 48 seconds. The value of this incoming "test case" score is 48 for the summed RT. We would then use Table 9 (Step 1) to determine the percentile for the raw score; a raw score of 48 yields the percentile is 83%. Next, the percentile standing is multiplied by the number of ranks (235), 235 * .83 ~ 195.05. A rank of 195 is then entered into Table 10 (Step 2) to find the "adjusted" T-score value for the incoming test case. Using this table, a rank of 195 would produce an adjusted T-score of 62.3.

The final step should be to compare the test case's adjusted T-score of 62.3 to the adjusted T-scores for the normal comparison group. In Table 8 (Step 3), we see that scores larger 62.3 occur in approximately 11% of the normal comparison group. This score is right on the boundary of a possible threshold or cut point for rejection (10 % threshold). Setting this threshold determines the sensitivity, and how many rejections will occur – setting this threshold to high (less than 10%) allows too many false negatives (e.g., test case is normal in performance when it truly comes from a non-normal population). Setting this threshold too low (greater than

10%) can allow too many false positives (e.g., saying a test case is not normal when it is not when the test case truly is normal). Selection of this threshold should be based on the practical costs of making an error in prediction. Which errors are of a greater consequence should be based on a clinically determined assessment of the degree of impact of those decision errors on the patient, and the treatment center. For our purposes of this experiment, a clinical group is used to assess the rejection rate with a known clinical population (e.g., dementia).

Table 11 displays demographics and the adjusted T-scores for a small, but known clinical population from validation group 2. The column titled "T-score" contains the clinical group's adjusted T-score values. Using an approximate 10% threshold, produces a threshold T-score of around 62. T-scores less than 62 would fail to be rejected as non-normal. On the basis of this scoring criterion, the 11 cases in the clinical group depicted in Table 11 produce a rejection rate of about 82%.

Checking Residualization and Rescaling T-scores for Ranked Summed RT

One check on the validity of the residualization and the subsequent rescaling to T-scores for the ranked summed RT, is to look at the correlations of the final T-score with the original unranked score, the ranked score, and background covariates. The transformed T-scores for the normal group should NOT be correlated with any of the background covariates but should still be highly correlated with both the accuracy and the RT either ranked or unranked. Table 12 demonstrates that this is true for the transformed T-scores. Essentially, adjusting the ranked summed RT scores on the background covariates worked very well ($r \sim 0$ for all covariates), with very little loss in the relationship with the original ranked summed RT ($r = .920$).

Lastly, one question arises as to whether the T-scores are now normally distributed after the residualization and T-score transformation. Figure 9 is a quantile-quantile plot (q-q) plot of

the theoretical quantiles of the normal distribution against the empirical quantiles of the adjusted T-scores for the normal group. If these two distributions overlap perfectly, the dotted line will lie completely along the straight reference line. Figure 9 indicates that within the center portion of the distribution that the data is very nearly normal, but that the tails of the data distribution are "heavy" or that too many cases are in the tails of the data distribution. This is to be expected since there are floor effects on the response latency (most people respond quickly to most items) and very poor performance happens infrequently but frequently enough to cause heavy tails on the high end of the RT data distribution. These features of the distribution of 235 normal comparison group make the case even more salient that the percentile or rank standing information of the population (N=235) is very important for being able to make inferences about incoming test scores. Making assumptions regarding normality in order to get percentile standing would be misleading in the case of skewed or heavy tailed distributions.

Tables 13, 14, and 15 are the percentile tables of the three covariate variables AGE, EDUC, and FSIQ. These tables are used to calculate a percentile score for a test case covariate, when comparing the test case's covariate score to the normal comparison group's covariate score. For example, using Table 15, for a test FSIQ score of 109, we would take the midpoint of the percentiles where FSIQ equals 109 in the normal comparison group (upper and lower value). Table 15 gives a value of 62 percentile for 109 and 61 percentile for 109. For an incoming score of 109 we would have a percentile of 61.5. That is, 61.5 percentile of the covariate scores in the comparison group had scores on FSIQ equal to or less than the incoming test case's FSIQ score. To convert this to a rank standing we would take .6125*235 to produce a rank value of 144. That is, 114 cases in the normal comparison group had a value equal to or smaller that the test case's covariate score.

Tables 13, 14, and 15 allow adjustments to be made on the actual values of the covariates whenever a test case has the full information on the covariates. Given values on AGE, EDUC, and FSIQ (with rank equal to 118 for gender, the average rank effect for gender), a clinician can use the canonical correlation and regression coefficients to produce an adjusted T-score for comparison to the normal comparison group. This would be better than using only the average rank information on the covariate adjustment procedure (average rank=118).

Preliminary sensitivity analyses showed that whenever extreme values on the covariates were present (e.g., FSIQ=130, EDUC=20, and AGE=75) that the adjustment itself can result in up to a 3 percentile points change in the calculated T-score. This adjustment would be a deciding factor whenever performance is near a threshold decision boundary (e.g., T-score of 62). This adjustment could be enough to cause a rejection in favor of non-normal decision for a test case that does not exceed the threshold.

CHAPTER 19

EXPERIMENT 2 DISCUSSION

As expected and shown in Experiment 1, the distribution of BNT-L scores in a sample of highly functioning adults was negatively skewed for both accuracy and latency to respond. Tests with this distribution are known to be highly discriminative at scores reflecting more severe impairment rather than scores reflecting mild or even moderate impairment. The purpose of Experiment 2 was to use the BNT-L to determine if latency measures increased the discriminative power to detect impairment in independent community-dwelling individuals who were referred for neuropsychological assessment with two classes of potential brain insult, mild traumatic brain injury and a form dementia or mild cognitive impairment.

Assessing reaction time offers important information about basic speed of processing which is an underlying mechanism that has been shown to mediate differences in cognitive functioning and can even affect a person's longevity (Deary & Der, 2005). Basic speed of processing appears to distinguish between impaired groups and normal controls. Felmingham et al. (2004) found group effects on more complex standard neuropsychological measures were removed in a sample of brain injured groups when basic speed of processing, as assessed by simple and choice reaction time, was controlled for statistically.

Basic speed of processing tasks are often impaired when an individual sustains a head injury, and have been shown when the head injury is moderate to severe (Perbal, Couillet, Azouvi, & Pouthas, 2003, slower reaction times found when GSC < 8). Our study did not find significant differences between the normative sample and sample with potential head injuries. This could either be due to this sample had no cognitive impairment, or they have sufficiently recovered from their injuries that the BNT-L was not sensitive enough to detect. Our group had "suspected" residual brain injuries following a documented closed injury that occurred at least 18

months prior to the evaluation. Significant recovery can occur in the following months after a head injury. Felmingham, Baguley, and Green (2004) administered a simple reaction time task on a population with diffuse injuries and found decreased speed of processing that was significant during 1-5 months of recovery that greatly improved after this period of time. Therefore, the sensitivity of the BNT-L may be greater for more acute or post acute injuries when processing speed is most affected.

The BNT-L's was sensitive to discriminating normal from abnormal response latencies in a population referred for neuropsychological assessment to evaluate for mild cognitive impairment or form of dementia. This group was significantly different from both the first validation group with potential acquired brain injury and the optimally healthy normative sample in both speed of response and accuracy measures for items on the BNT-L. This group was similar to the first validation group in that they were highly functioning adults living independently in the community. In addition to differences in reason for referral, this group had several differences from potential brain injury group. The main difference between the validation groups is the age of the sample. The first group comprised a significantly larger age range and included younger individuals (ages 21 to 64) whereas the group with potential dementia comprised elderly individuals with a small range of ages (62-83). Perhaps the BNT-L is most sensitive to an aging population when decrements in naming abilities with age are detectable with measuring latencies in whole seconds.

Response styles were different for the population with a suspected dementia condition from the normative sample and the younger validation group. Obler and Albert (1985) and Nichols et al. (1985) had analogous findings after examining different response styles between younger and older age groups. As found in our validation samples, both these studies found older

people in their sample produced more circumlocutions and picture descriptions than younger people. Increased verbalizations during problem-solving a naming failure provides information regarding controlled processing for that individual (Stern et al., 1991) as well as search strategies and points of linguistic failure (Goodglass et al., 2001).

Additional information concerning reaction time or basic speed of processing is also useful information when assessing the integrity of one's neurocognitive abilities Deary & Der, 2005). Further research with latency to respond to naming tasks is warranted, especially with information to help detect specific areas of cognition that are affected. The BNT already provides a mechanism for determining the linguistic point of failure in the word finding process, however, differences in latencies can correlate with the type of information (conceptual, semantic, or phonological) that can be recovered, and this information to be useful in determining affected areas of cognition and subsequent remediation to help improve an individual's quality of life. An apparent limitation using these validation groups is the small sample size of each population. Other information was lacking such as the results or degree of impairment determined following the neuropsychological battery in which the BNT-L was incorporated, or if the whether the individual listed word-finding difficulties as a symptom or complaint.

Further validation of the BNT-L is needed with regards to acquired head injuries, particularly within specific points of recovery. The BNT-L's measurement of reaction time in addition to naming can be dually beneficial since simple reaction time is not a typical part of a neuropsychological battery and reaction time has been shown to affect performance on more complex measures.

In the end, it is hoped that more precise latencies not only provide a sensitive indicator of normal aging processing in naming ability that can be used as a comparison for higher functioning abnormal populations, but also as a useful gauge for assessing recovery of function in individuals with naming difficulties or anomia (Dunn et al., 1989).

Table 1

*Demographics of BNT-L Sample (*N = *235)*

Age	Education	Gender	Race	Handedness
18-89	7-20	45.5% Male	82.1% White	89.2% Right
M = 43	M = 13.82	54.5% Fem	11.1% African American	9.1% Left
			4.7% Hispanic	1.7% Ambidextrous
			2.2% Asian, Other	

Table 2. *Codes Used, as needed, for Qualitative Experience of Normative Sample*

RN	Related Name:	Incorrect response not specific or precise enough for the target word, e.g., boat for canoe; rope for noose; statue for sphinx; lock for latch
SS	Similar Sounds:	Phonemic pharaphasia or colloquialism. Articulated response has similar phonetic sound but is incorrect, e.g., trestle for trellis; abscuss for abacus.
BP	Bad Picture	When the picture was misperceived due to features of the drawing and not due to a visual perceptual dysfunction, e.g., easel for tripod.
SW	Semantic Word	Incorrect response that is within the same semantic category as the target word, e.g., lattice for trellis; pharaoh for sphinx.
SQ	Stimulus Cue	Cue given to clarify the picture when the picture is misperceived or the person lacks recognition of the picture ("I don't know what that is"), e.g., incorrect response of protractor for "tripod" cue given from standard BNT protocol, "photographers or surveyors use it."
PQ	Phonemic Cue	Given after 30 seconds if the target response is not provided. Note: This cue typically resolves a TOT state.
DKW	Don't Know Word	Unfamiliarity with the word caused a nonresponse rather than inadequate word finding; the word does not exist in their lexical repertoire.
VW	Vocabulary Word	Picture stimulus was assessing vocabulary rather than word-finding. The target word was recognized but was not the word they were trying to retrieve or was not a word they would retrieve independently with or without a cue.
V	Verbalization	Paralinguistic activity where an individual verbally describes the item as a problem-solving strategy

Table 3. *Frequency of Codes on Individual Items*

Item ***(no code)***	***RN*** ***1***	***SS*** ***2***	***BP*** ***3***	***SW*** ***4***	***SQ*** ***5***	***PQ*** ***6***	***DKW*** ***7***	***VW*** ***8***	***V*** ***9***	***TOT?***
Tree (235)										
Bed (235)										
Pencil (235)										
House (235)										
Whistle (234)				1						
Scissors (235)										
Comb (235)										
Flower (234)			1		1					
Saw (235)										
Toothbrush (232)			3							
Helicopter (227)		1		7 (plane)						
Broom (235)										

Table 3 *continued*.

Item ***(no code)***	***RN*** ***1***	***SS*** ***2***	***BP*** ***3***	***SW*** ***4***	***SQ*** ***5***	***PQ*** ***6***	***DKW*** ***7***	***VW*** ***8***	***V*** ***9***	***TOT?***
Octopus (225)	1		2	6	1	1				1
Mushroom (234)	1									
Hanger (233)				2						
Wheelchair (235)										
Camel (233)				2						1
Mask (232)			1	1					1	
Pretzel (231)		1	2		2	1			1	1
Bench (234)						1			1	
Racquet (234)			1							
Snail (232)				1	2	2	2	2		
Volcano (230)					1	4			1	1

Table 3 *continued*.

Seahorse (218)				2	2	10	5	5	6	5
Dart (222)	1		3	4	4	2			1	2
Canoe (223)	8			3	1	3	1	1		2
Globe (221)	3			7		2			3	3
Wreath (233)			1	1	2	1	1	1		
Beaver (188)	2	1	3	31	3	16			7	13
Harmonica (231)			1	1	1	2				1
Rhinoceros (223)				9 (hippo)		4			1	4
Acorn (225)	3			2		5			3	2
Igloo (226)	2		1		2	7	2	2		5
Stilts (217)	2	1		2	7	16	7	8	1	6
Domino (228)				5 (dice)	1	1				1
Cactus (232)						2				2

Table 3 *continued*.

Escalator (227)	1			4	1	2				2
Harp (231)				1		3		1	1	3
Hammock (208)	2	1	2		3	21	2	2	8	17
Knocker (213)	2	2	4	2	3	7	3	6	4	4
Pelican (189)	5		3	28	8	23	7	12	2	15
Stethoscop (215)		7			4	15	4	4	6	12
Pyramid (227)				2	3	8	3	3	2	5
Muzzle (173)	5	3	17	12	15	35	10	10	10	25
Unicorn (220)		3	1	3	3	10	3	4		7
Funnel (214)	2		1	3	1	14			9	15
Accordion (179)	5	2		8	5	44	5	5	21	38
Noose (198)	17	4		1	16	24	15	15	1	8

Table 3 *continued*.

Asparagus (204)	1	1	15 (branch)	8	17	9			3	9
Compass (122)		1		24 (protract)	14	63	39	48	12	37
Latch (168)	47 (lock)	2	3	4	10	34	10	16	3	21
Tripod (147)	4	3	45 (compass)	4	76	28	13	14	7	16
Scroll (201)	3	2		7	8	24	7	8	3	16
Tongs (190)	16	15 (prongs; thongs)		2	2	20	7	8	3	19
Sphinx (158)		1	1	13	30	69	30	34	17	34
Yoke (109)	1	1		4	100	121	90	92	7	25
Trellis (107)	1	17		30	72	108	71	65	7	25
Palette (144)	14	1		17	33	75	31	35	13	43
Protractor (109)		2	1	69	47	97	40	50	10	34

Table 3 *completed*.

Abacus (115)		9	2	4	81	10 6	80	82	14	22

Table 4a. *Common Code Sequence and Corresponding Frequency of Items (Table of Code Sequences Continues with Table 4b)*

RN/SQ 1,5	***RN/PQ 1,6***	***RN/V 2, 9***	***BP/SW 3,4***	***BP/SQ 3,5***	***SW/PQ 4,6***	***PQ/V 6,9***	***SQ/PQ 5,6***	***SQ/PQ/DKW/VW 5,6,7,8***	***RN/SQ/PQ/DKW/VW 1,5,6,7,8***	***SQ/PQ/DKW/VW/V 5,6,7,8,9***
Dart (1)	Canoe (1)	Beaver (1)	Dart (1)	Flower (1)	Seahorse (1)	Bench (1)	Dart (1)	Snail (2)	Canoe (1)	Seahorse (1)
	Beaver (1)	Stethosc (1)	Beaver (1)	Octopus (1)	Globe (1)	Volcano (1)	Asparagu (2)	Seahorse (4)	Pelican (1)	Stilts (1)
	Igloo (1)	Accordi (1)	Pelican (1)	Pretzel (2)	Beaver (1)	Beaver (3)	Tripod (10)	Wreath (1)	Latch (3)	Stethoscope (1)
	Hamm (2)		Asparagus (1)	Dart (3)	Rhino(2)	Acorn (2)	Protractor (5)	Igloo (2)	Tripod (1)	Pyramid (1)
	Pelica(1)			Wreath (1)	Acorn (1)	Igloo (1)		Stilts (6)	Tongs (1)	Compass (3)
	Accor (2)			Beaver (1)	Pyramid (1)	Cactus (1)		Hammock (2)	Sphinx (2)	Latch (1)
	Latch (2)			Harmonica (1)	Unicorn (1)	Hammoc (5)		Knocker (3)	Trellis (1)	Tripod (1)
	Tongs (6)			Igloo (1)	Accordion (3)	Pelican (2)		Pelican (5)		Sphinx (5)
				Funnel (1)	Asparagus (2)	Stethosc (1)		Stethoscope (4)		Trellis (4)
				Asparagus (8)	Protractor (6)	Pyramid (1)		Pyramid (1)		Protractor (2)
						Funnel (8)		Unicorn (2)		
						Accordio n (12)		Accordion (5)		

Table 4a *continued*.

.

RN/SQ 1,5	***RN/PQ 1,6***	***RN/V 2, 9***	***BP/SW 3,4***	***BP/SQ 3,5***	***SW/PQ 4,6***	***PQ/V 6,9***	***SQ/PQ 5,6***	***SQ/PQ/DKW/VW 5,6,7,8***	***RN/SQ/PQ/DKW/VW 1,5,6,7,8***	***SQ/PQ/DKW/VW/V 5,6,7,8,9***
						Asparag (2)		Compass (35)		
						Compass (8)		Latch (6)		
						Latch (2)		Tripod (10)		
						Tripod (2)		Sphinx (20)		
						Tongs (2)		Trellis (53)		
						Sphinx (10)		Protractor (38)		
						Trellis (3)				
						Protracto (4)				

Table 4b. *Common Code Sequence and Corresponding Frequency of Items (Table of Codes Continued from Table 4a)*

SW, PQ, V 4,6,9	***PQ/VW/V 6,8,9***	***BP/SW/SQ /PQ 3,4,5,6***	***SW/SQ/PQ /DKW/VW 4,5,6,7,8***	***PQ/VW 6,8***	***SS.PQ 2,6***
Beaver (1)	Stilts (1)	Pelican (1)	Pelican (1)	Stilts (1)	
Acorn (1)	Harp (1)	Tripod (1)	Pyramid (1)	Compass (6)	Trellis (5)
Accordion (2)	Knocker (1)		Unicorn (2)	Latch (2)	
Sphinx (3)	Pelican (1)		Tripod (1)	Tongs (2)	
			Sphinx (4)	Sphinx (1)	
			Trellis (11)	Trellis (5)	
				Protractor (1)	

Table 5. *Spearman's Rank Correlation on All 60 Items with Potentially Confounding Variables*

Potential Confounding Variable	**Spearman's Rank Correlation (*r*)**	**Level of Significance**
Age	.16	P = .001
Gender	.13	P = .001
Education	-.22	P = .001
FSIQ	-.41	P = .001
Accuracy	-0.88	P = .001

Table 6. *Maximum Likelihood Confirmatory Factor Analysis based on Spearman Ranks*

Uniquenesses:

HELICOPT	HANGER	RACQUET	SNAIL	SEAHORSE	DART25	RHINO	ACORN	IGLOO	CACTUS
0.843	0.955	0.928	0.786	0.909	0.899	0.715	0.818	0.467	0.905

ESCALATR	HARP	PYRAMID	FUNNEL	ASPARAGUS
0.872	0.898	0.717	0.970	0.939

Loadings:

Number	Item	Loading
11	HELICOPT	0.396
15	HANGER	0.213
21	RACQUET	0.267
22	SNAIL	0.462
24	SEAHORSE	0.302
25	DART25	0.319
31	RHINO	0.534
32	ACORN	0.426
33	IGLOO	**0.730**
36	CACTUS	0.309
37	ESCALATR	0.358
38	HARP	0.319
43	PYRAMID	0.532
46	FUNNEL	**0.172**
49	ASPARAGU	0.247

Fit Indices For Spearman Correlation Matrix:

Model Chi-square = 130.42, Df=90, Pr(>Chi-sq) = 0.003462
Chi-square (null model) = 438.75, Df = 105
Goodness-of-fit index = 0.9285
Adjusted goodness-of-fit index = 0.90466
RMSEA index = 0.043811, 90% CI: (0.025742, 0.059633)
Bentler-Bonnett NFI = 0.70274
Tucker-Lewis NNFI = 0.8587
Bentler CFI = 0.87888
BIC = -360.94

Fit Indices For Pearson Correlation Matrix:

RMSEA index = 0.12524 90% CI: (0.1133, 0.13745)
Goodness-of-fit index = 0.81351
Adjusted goodness-of-fit index = 0.75134
BIC = -71.016

Reliabilities:

		Omega (Spearman)	Omega (Pearson R)	Alpha (Spearman)
SS loadings	2.379	0.7120558	.679	.677
Proportion Var	0.159			

Table 7. *Bootstrap Means and Standard Errors for Predictive Equation Coefficients*

*Wx*Xrrt = Intercept + Slope*[W1 * Y1age + W2 * Y2educ + W3 * Y3gender + W4 * Y4fsiq]*

Coefficient * Variable Name	**Bootstrap Mean Of Coefficients**	**Standard Error Of Coefficients**
Wx * Xrrt (Xrrt - Xcomposite)	Wx = .000006	.000027865
C * Ycomposite (C – canonical correlation)	C = .40	.001838855
Intercept	Intercept =.00088	.004136289
W1 * Y1age ----------|	W1 = .0000115	.000006989
W2 * Y2educ ------- | ----- Ycomposite	W2 = .00000043	.000008009
W3 * Y3gender ------ |	W3 = .00000024	.000006416
W4 * Y4fsiq ----------|	W4 = .0000056	.000028621

Table 8. *T-Score Percentiles for Normal Comparison Group (T < 62 ~ 90%ile) (Step 3)*

100%	99%	98%	97%	96%	95%	94%	93%
68.10980	67.57121	66.27967	65.86404	64.91230	64.60066	63.72627	63.44229
92%	91%	90%	89%	88%	87%	86%	85%
63.25547	62.98798	62.77828	62.41893	62.14986	61.98721	61.90003	61.72492
84%	83%	82%	81%	80%	79%	78%	77%
61.56363	61.16211	60.64516	60.58735	60.17305	59.66626	59.45502	59.23899
76%	75%	74%	73%	72%	71%	70%	69%
59.15547	58.63118	58.44017	58.26840	58.06080	57.39724	56.97325	56.79467
68%	67%	66%	65%	64%	63%	62%	61%
56.59849	56.41739	55.65955	55.26360	55.07061	54.69019	54.42247	53.80310
60%	59%	58%	57%	56%	55%	54%	53%
53.65067	53.52981	53.08356	52.64004	52.27936	51.59009	51.20941	50.95366
52%	51%	50%	49%	48%	47%	46%	45%
50.67579	50.42394	50.10843	49.93735	49.86919	49.76211	49.48600	49.00486
44%	43%	42%	41%	40%	39%	38%	37%
48.55826	48.27856	47.86844	47.63542	47.04000	46.86435	46.50787	46.40831
36%	35%	34%	33%	32%	31%	30%	29%
45.98084	45.80209	45.59782	45.48125	45.34344	43.61875	43.51808	42.97382
28%	27%	26%	25%	24%	23%	22%	21%
42.66598	42.14182	41.21859	40.98567	40.62864	40.48149	40.33152	40.08748
20%	19%	18%	17%	16%	15%	14%	13%
40.02860	39.72510	39.56320	39.44890	39.29564	38.85858	38.33456	37.73581
12%	11%	10%	9%	8%	7%	6%	5%
37.40821	37.26339	37.10899	35.95101	34.78564	34.42073	33.79731	33.68726
4%	3%	2%	1%	0%			
32.80635	32.09601	31.17677	30.04489	29.28432			

Table 9. *Percentiles of Summed Reaction Times and Descriptive Statistics (Step 1)*

100%	99%	98%	97%	96%	95%	94%	93%	92%	91%	90%	89%	88%	87%	86%	85%	84%	83%	82%
249	163	127	94	89	84	77	73	69	64	61	57	55	52	52	51	50	48	47
81%	80%	79%	78%	77%	76%	75%	74%	73%	72%	71%	70%	69%	68%	67%	66%	65%	64%	63%
46	46	45	42	40	38	36	36	34	33	32	32	30	29	29	29	27	26	26
62%	61%	60%	59%	58%	57%	56%	55%	54%	53%	52%	51%	50%	49%	48%	47%	46%	45%	44%
24	24	23	23	23	22	22	21	21	21	21	20	20	20	20	19	19	19	19
43%	42%	41%	40%	39%	38%	37%	36%	35%	34%	33%	32%	31%	30%	29%	28%	27%	26%	25%
19	18	18	18	18	18	18	17	17	17	17	17	17	17	16	16	16	16	16
24%	23%	22%	21%	20%	19%	18%	17%	16%	15%	14%	13%	12%	11%	10%	9%	8%	7%	6%
16	16	16	16	16	16	16	15	15	15	15	15	15	15	15	15	15	15	15
5%	4%	3%	2%	1%	0%
15	15	15	15	15	15

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	N
15.00	16.00	20.00	32.39	36.50	249.00	30.56	235

Table 10. *Converting Unadjusted Rank to Adjusted T-Scores (Step 2)*

Rk	T-score	Rk	T-score	Rk	T-score	Rk	T-score	Rk	T-score
1	31.29069	51	39.28612	101	47.28155	151	55.27698	201	63.27242
2	31.45060	52	39.44603	102	47.44146	152	55.43689	202	63.43232
3	31.61051	53	39.60594	103	47.60137	153	55.59680	203	63.59223
4	31.77042	54	39.76585	104	47.76128	154	55.75671	204	63.75214
5	31.93033	55	39.92576	105	47.92119	155	55.91662	205	63.91205
6	32.09023	56	40.08567	106	48.08110	156	56.07653	206	64.07196
7	32.25014	57	40.24557	107	48.24101	157	56.23644	207	64.23187
8	32.41005	58	40.40548	108	48.40091	158	56.39634	208	64.39178
9	32.56996	59	40.56539	109	48.56082	159	56.55625	209	64.55168
10	32.72987	60	40.72530	110	48.72073	160	56.71616	210	64.71159
11	32.88978	61	40.88521	111	48.88064	161	56.87607	211	64.87150
12	33.04969	62	41.04512	112	49.04055	162	57.03598	212	65.03141
13	33.20959	63	41.20503	113	49.20046	163	57.19589	213	65.19132
14	33.36950	64	41.36493	114	49.36037	164	57.35580	214	65.35123
15	33.52941	65	41.52484	115	49.52027	165	57.51571	215	65.51114
16	33.68932	66	41.68475	116	49.68018	166	57.67561	216	65.67104
17	33.84923	67	41.84466	117	49.84009	167	57.83552	217	65.83095
18	34.00914	68	42.00457	118	50.00000	168	57.99543	218	65.99086
19	34.16905	69	42.16448	119	50.15991	169	58.15534	219	66.15077
20	34.32896	70	42.32439	120	50.31982	170	58.31525	220	66.31068
21	34.48886	71	42.48429	121	50.47973	171	58.47516	221	66.47059
22	34.64877	72	42.64420	122	50.63963	172	58.63507	222	66.63050
23	34.80868	73	42.80411	123	50.79954	173	58.79497	223	66.79041
24	34.96859	74	42.96402	124	50.95945	174	58.95488	224	66.95031
25	35.12850	75	43.12393	125	51.11936	175	59.11479	225	67.11022
26	35.28841	76	43.28384	126	51.27927	176	59.27470	226	67.27013
27	35.44832	77	43.44375	127	51.43918	177	59.43461	227	67.43004
28	35.60822	78	43.60366	128	51.59909	178	59.59452	228	67.58995
29	35.76813	79	43.76356	129	51.75899	179	59.75443	229	67.74986
30	35.92804	80	43.92347	130	51.91890	180	59.91433	230	67.90977
31	36.08795	81	44.08338	131	52.07881	181	60.07424	231	68.06967
32	36.24786	82	44.24329	132	52.23872	182	60.23415	232	68.22958
33	36.40777	83	44.40320	133	52.39863	183	60.39406	233	68.38949
34	36.56768	84	44.56311	134	52.55854	184	60.55397	234	68.54940
35	36.72758	85	44.72302	135	52.71845	185	60.71388	235	68.70931
36	36.88749	86	44.88292	136	52.87836	186	60.87379		
37	37.04740	87	45.04283	137	53.03826	187	61.03369		
38	37.20731	88	45.20274	138	53.19817	188	61.19360		
Cont below		*Cont below*		*Cont below*		*Cont below*			

Table 10 *column continuation.*

Cont from above		*Cont from above*		*Cont from above*		*Cont from above*	
39	37.36722	89	45.36265	139	53.35808	189	61.35351
40	37.52713	90	45.52256	140	53.51799	190	61.51342
41	37.68704	91	45.68247	141	53.67790	191	61.67333
42	37.84694	92	45.84238	142	53.83781	192	61.83324
43	38.00685	93	46.00228	143	53.99772	193	61.99315
44	38.16676	94	46.16219	144	54.15762	194	62.15306
45	38.32667	95	46.32210	145	54.31753	195	62.31296
46	38.48658	96	46.48201	146	54.47744	196	62.47287
47	38.64649	97	46.64192	147	54.63735	197	62.63278
48	38.80640	98	46.80183	148	54.79726	198	62.79269
49	38.96631	99	46.96174	149	54.95717	199	62.95260
50	39.12621	100	47.12164	150	55.11708	200	63.11251

Table 11. *Demographics and Results for Validation Group 2 (Referral: MCI or Dementia)*

ID	AGE	EDU	GENDER	HVOT	BNTL	RANK	T-score	HAND	TOT	Accuracy
1	76	13	F	19.5	48	195	62.31	R	1	15
2	69	16	M	18.0	175	234	68.55	R	NA	12
3	68	15	F	23.5	143	233	68.34	R	1	14
4	79	12	F	21.0	306	235	68.71	R	NA	11
5	62	18	F	25.5	17	85	**44.72**	R	0	15
6	83	10	F	28.0	30	162	**57.04**	R	NA	15
7	70	16	F	25.0	118	229	67.75	R	NA	14
8	77	13	F	24.0	173	234	68.43	R	NA	14
9	69	12	M	24.0	175	235	68.55	R	NA	14
10	73	12	M	12.0	186	235	68.60	R	8	12
11	80	16	M	14.0	246	235	68.71	R	2	12

Rejection Rate = (1-2/11) = 82% (Bolded Entries Not Rejected As Normal)

Mean Age = 73.27, SD = 6.29; Mean Accuracy = 13.45, SD = 1.4

Table 12. *Validity Correlations: Spearman's Rank Correlation between Adjusted T-scores and Original Scores; Ranked Scores on Summed RT; and Background Covariates*

	Spearman's Rank Correlation (*r*) Final Adjusted T-scores for Normal Group
Un-Ranked Summed RT	0.5940
Ranked Summed RT	0.9270
Ranked Age	-0.0089
Ranked Gender	0.0230
Ranked Education	0.0270
Ranked FSIQ	-0.0088
Ranked Accuracy	-0.6653

Table 13. *Percentiles of Age and Descriptive Statistics*

100%	99%	98%	97%	96%	95%	94%	93%	92%	91%	90%	89%	88%	87%	86%	85%
89	85	84	80	80	79	79	77	76	75	75	74	73	72	72	70
84%	83%	82%	81%	80%	79%	78%	77%	76%	75%	74%	73%	72%	71%	70%	69%
69	69	68	67	66	65	64	63	62	60	59	58	58	56	56	55
68%	67%	66%	65%	64%	63%	62%	61%	60%	59%	58%	57%	56%	55%	54%	53%
54	54	54	53	51	50	50	49	48	47	47	46	45	44	44	43
52%	51%	50%	49%	48%	47%	46%	45%	44%	43%	42%	41%	40%	39%	38%	37%
42	40	40	39	39	38	38	36	36	35	34	33	31	29	28	27
36%	35%	34%	33%	32%	31%	30%	29%	28%	27%	26%	25%	24%	23%	22%	21%
27	26	26	25	25	25	24	24	24	23	23	23	22	22	22	22
20%	19%	18%	17%	16%	15%	14%	13%	12%	11%	10%	9%	8%	7%	6%	5%
22	21	21	21	21	21	21	21	20	20	20	20	20	20	19	19
4%	3%	2%	1%	0%											
19	19	19	18	18											

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	N
18.00	23.00	40.00	43.31	60.50	89.00	20.93	235

Table 14. *Percentiles of Education and Descriptive Statistics*

100%	99%	98%	97%	96%	95%	94%	93%	92%	91%	90%	89%	88%	87%	86%	85%
20	19	18	18	18	18	18	18	17	17	16	16	16	16	16	16
84%	83%	82%	81%	80%	79%	78%	77%	76%	75%	74%	73%	72%	71%	70%	69%
16	16	16	16	16	16	16	16	16	15	15	15	15	15	15	15
68%	67%	66%	65%	64%	63%	62%	61%	60%	59%	58%	57%	56%	55%	54%	53%
15	15	15	15	15	15	15	14	14	14	14	14	14	14	14	14
52%	51%	50%	49%	48%	47%	46%	45%	44%	43%	42%	41%	40%	39%	38%	37%
14	14	14	14	14	14	14	13	13	13	13	13	13	13	13	13
36%	35%	34%	33%	32%	31%	30%	29%	28%	27%	26%	25%	24%	23%	22%	21%
12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
20%	19%	18%	17%	16%	15%	14%	13%	12%	11%	10%	9%	8%	7%	6%	5%
12	12	12	12	12	12	12	12	12	12	12	12	12	11	10	10
4%	3%	2%	1%	0%											
10	10	9	9	7											

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	N
7.00	12.00	14.00	13.82	15.00	20.00	2.28	235

Table 15. *Percentiles of FSIQ and Descriptive Statistics*

100%	99%	98%	97%	96%	95%	94%	93%	92%	91%	90%	89%	88%	87%	86%	85%
125	124	122	121	121	120	119	119	118	118	118	117	117	117	116	116
84%	83%	82%	81%	80%	79%	78%	77%	76%	75%	74%	73%	72%	71%	70%	69%
116	115	115	115	115	115	115	114	114	114	113	112	112	112	111	111
68%	67%	66%	65%	64%	63%	62%	61%	60%	59%	58%	57%	56%	55%	54%	53%
110	110	110	110	110	110	109	109	108	108	108	107	107	106	106	105
52%	51%	50%	49%	48%	47%	46%	45%	44%	43%	42%	41%	40%	39%	38%	37%
105	104	104	103	103	103	103	102	102	102	102	102	101	101	101	101
36%	35%	34%	33%	32%	31%	30%	29%	28%	27%	26%	25%	24%	23%	22%	21%
100	100	100	99	99	99	98	98	98	97	97	97	96	96	96	96
20%	19%	18%	17%	16%	15%	14%	13%	12%	11%	10%	9%	8%	7%	6%	5%
95	95	95	94	93	93	92	92	92	91	90	89	89	88	87	85
4%	3%	2%	1%	0%											
84	83	82	80	77											

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	N
77.0	97.0	104.0	104.3	114.0	125.0	10.68	235

Figure 1. Serial stage model of naming.

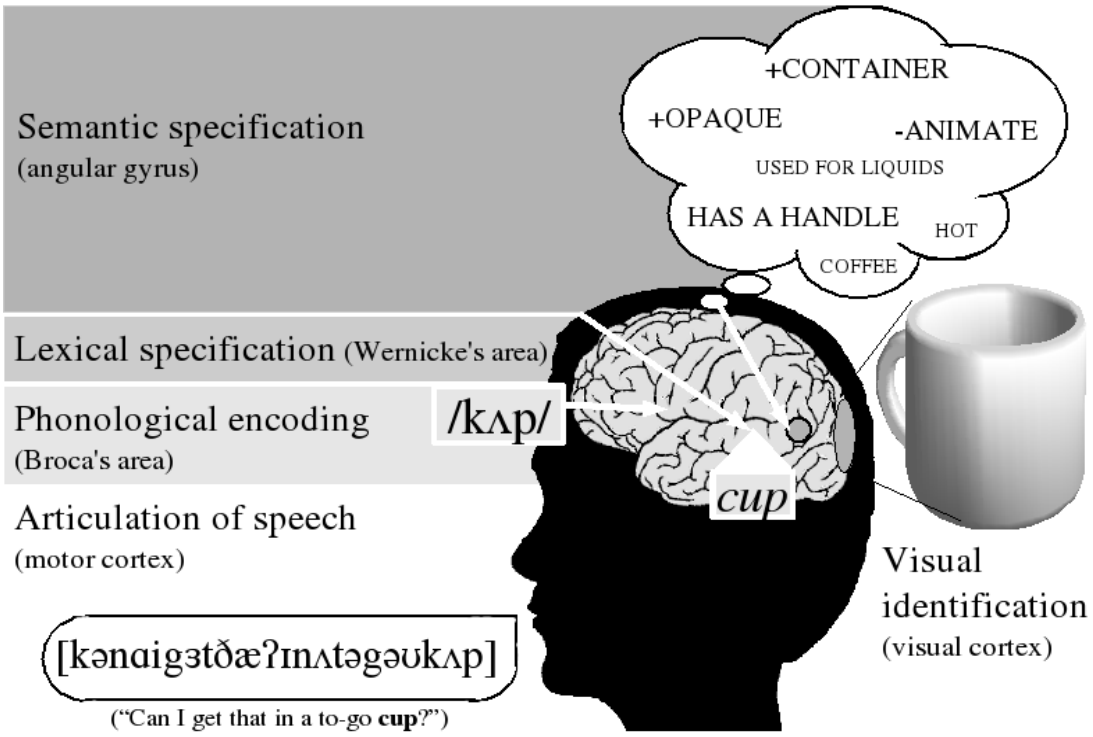


Figure 2. Histograms of summed RT score and background covariates.

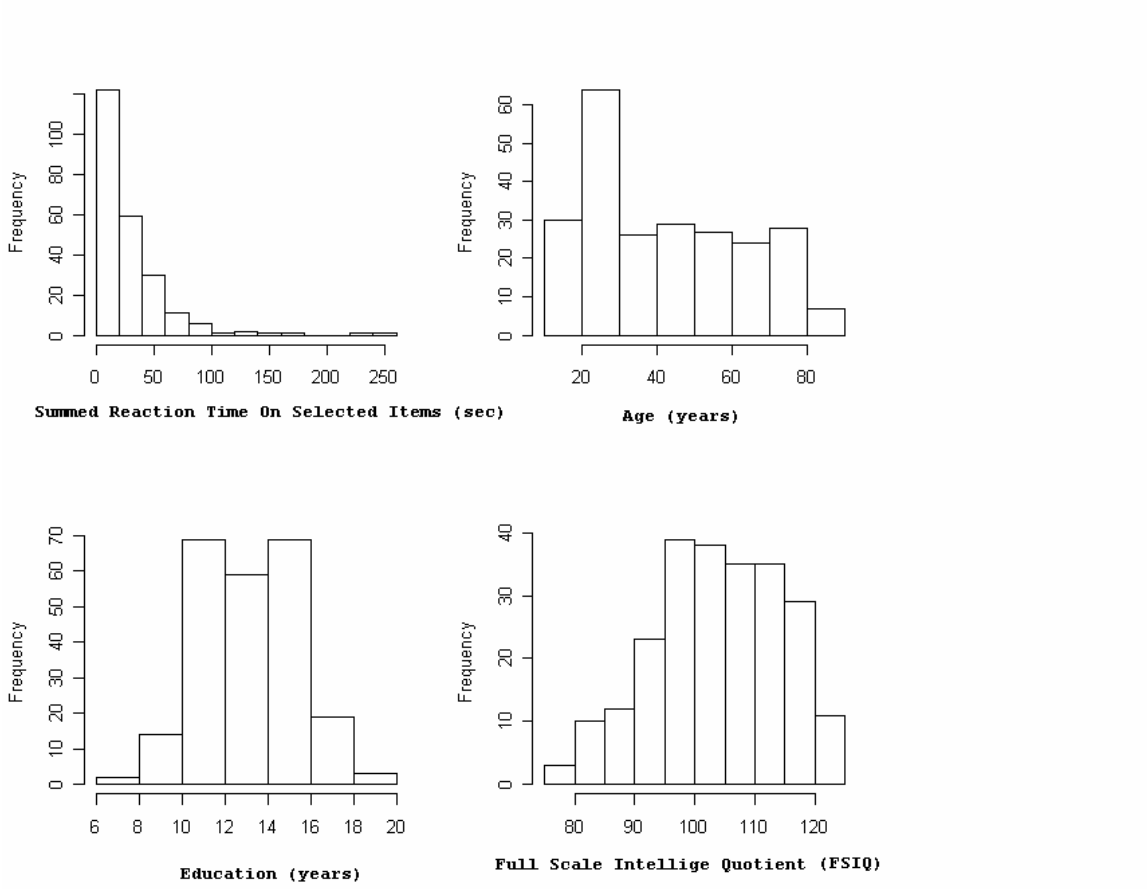


Figure 3. Sum reaction time plotted against summed accuracy for selected 15 items.

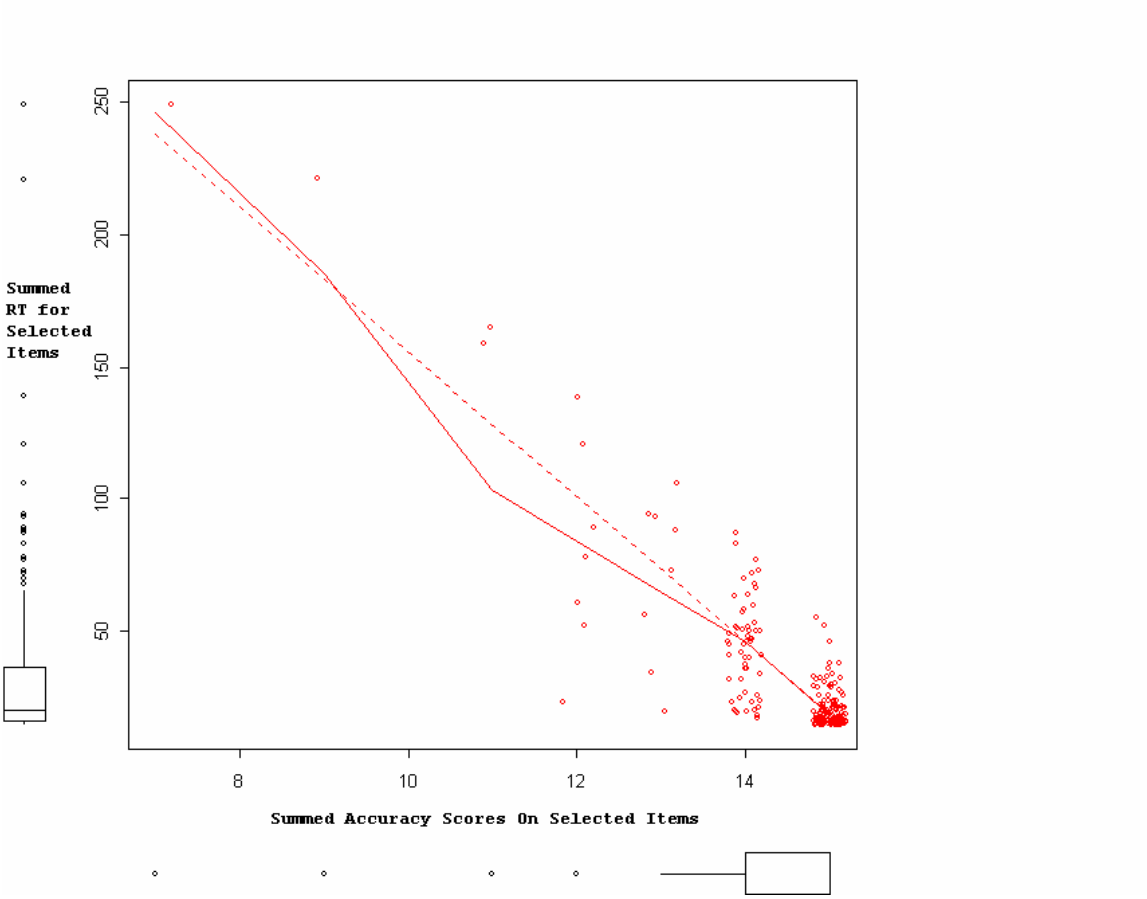


Figure 4. Ranks of summed reaction times plotted against ranks of summed accuracy for selected 15 items.

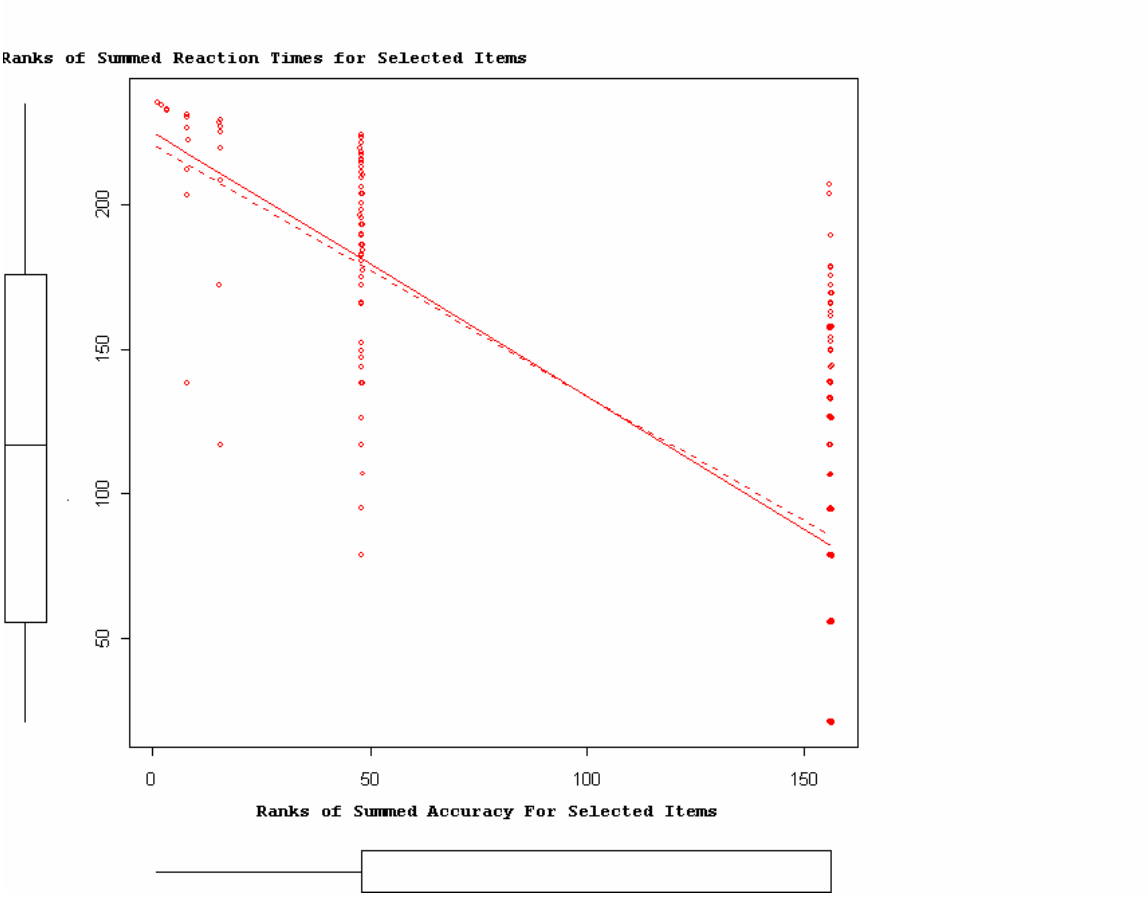


Figure 5. Confirmatory factor analysis loadings ordered by item.

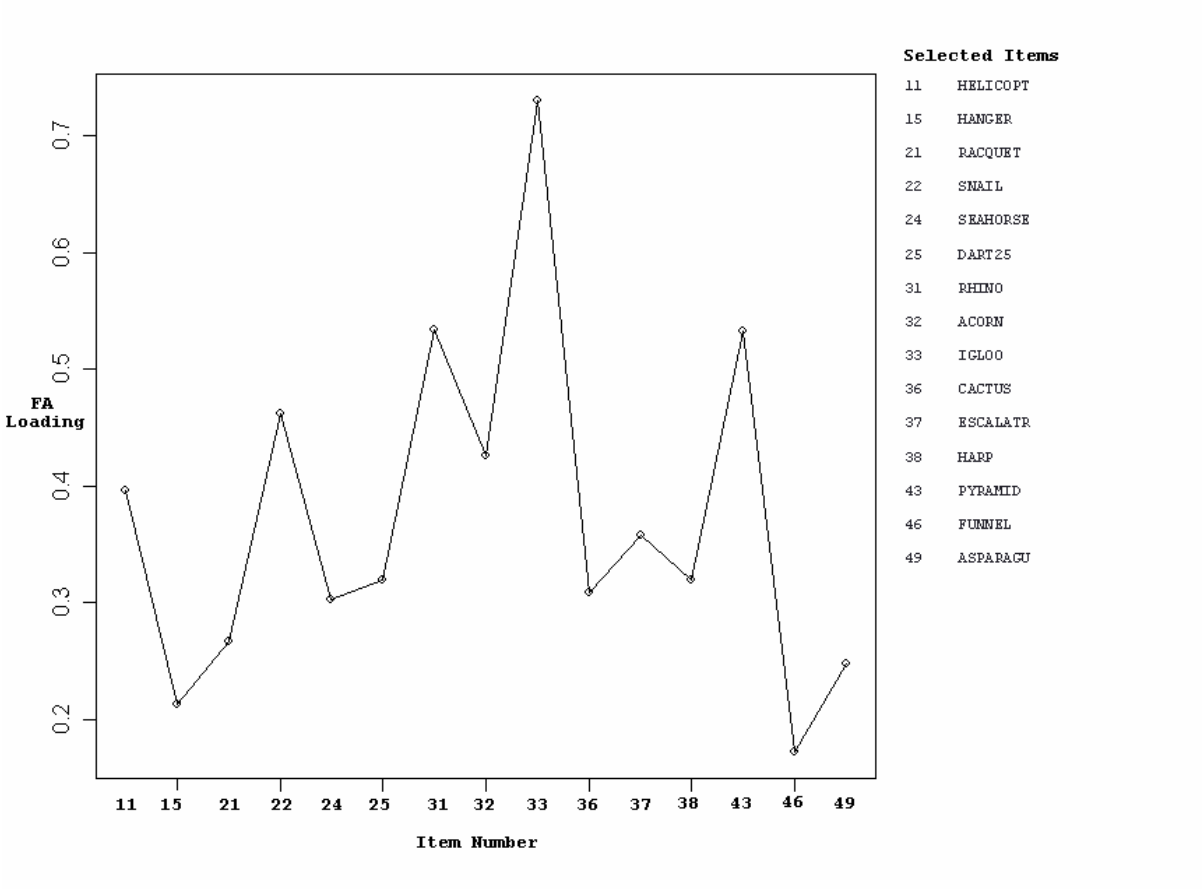


Figure 6. Accuracy of all items ordered by item number.

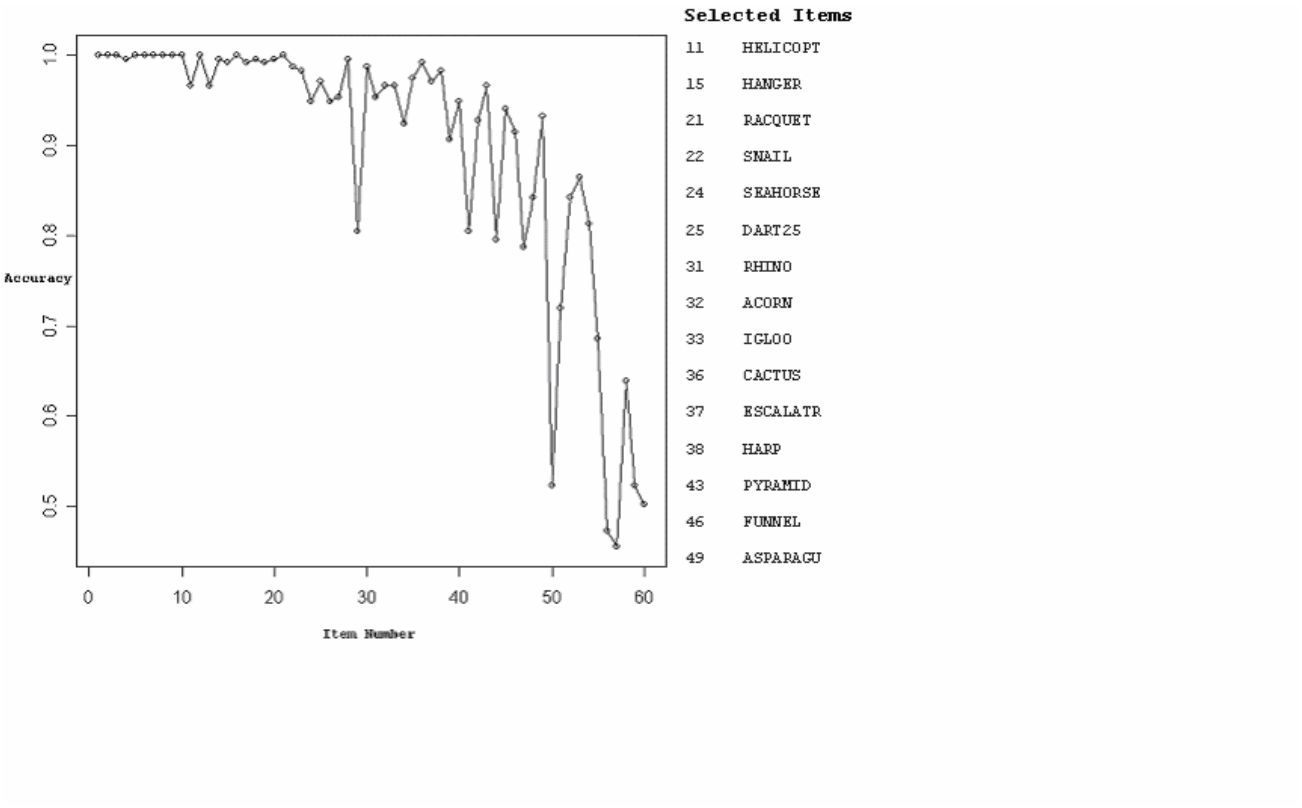


Figure 7. Item information (loading divided by error) ordered by item selected.

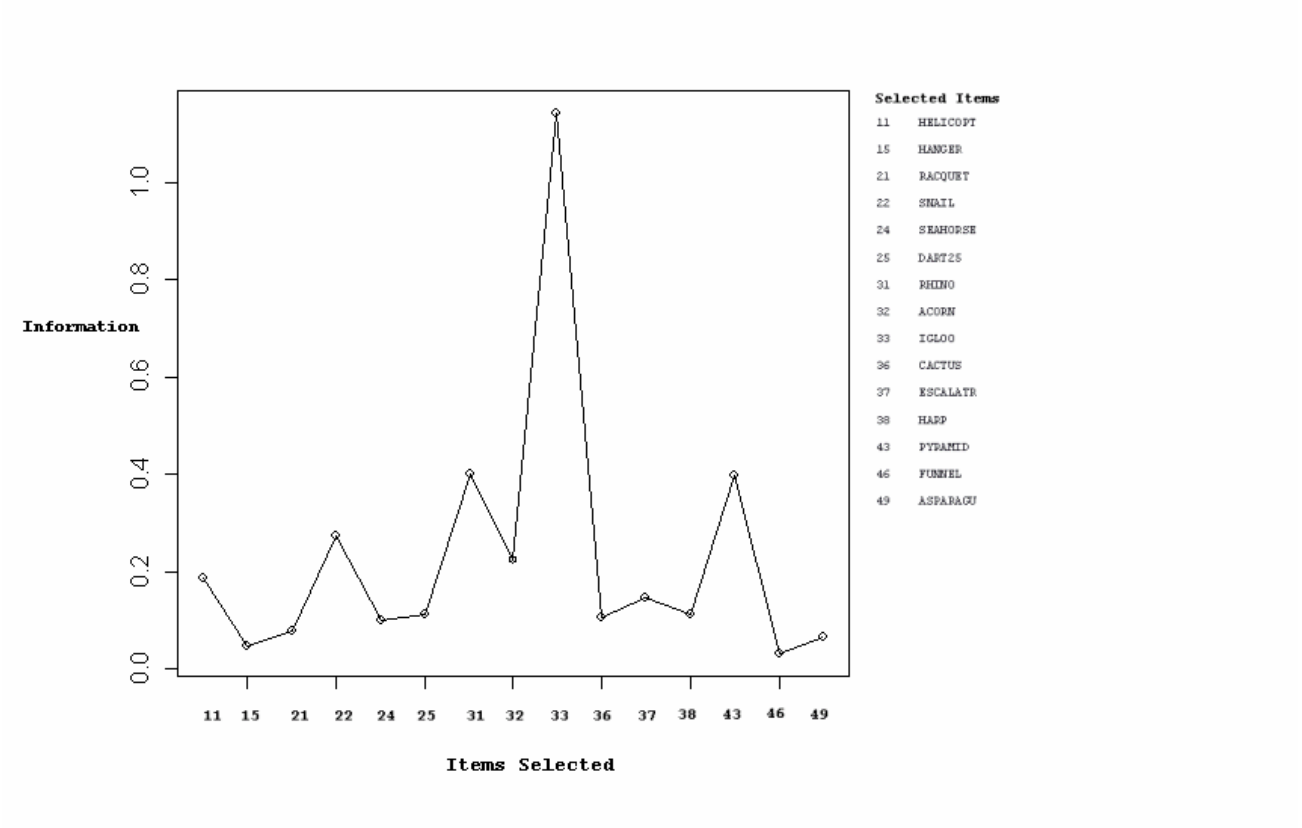


Figure 8. Reaction time item response curves ordered by ability levels.
on block one

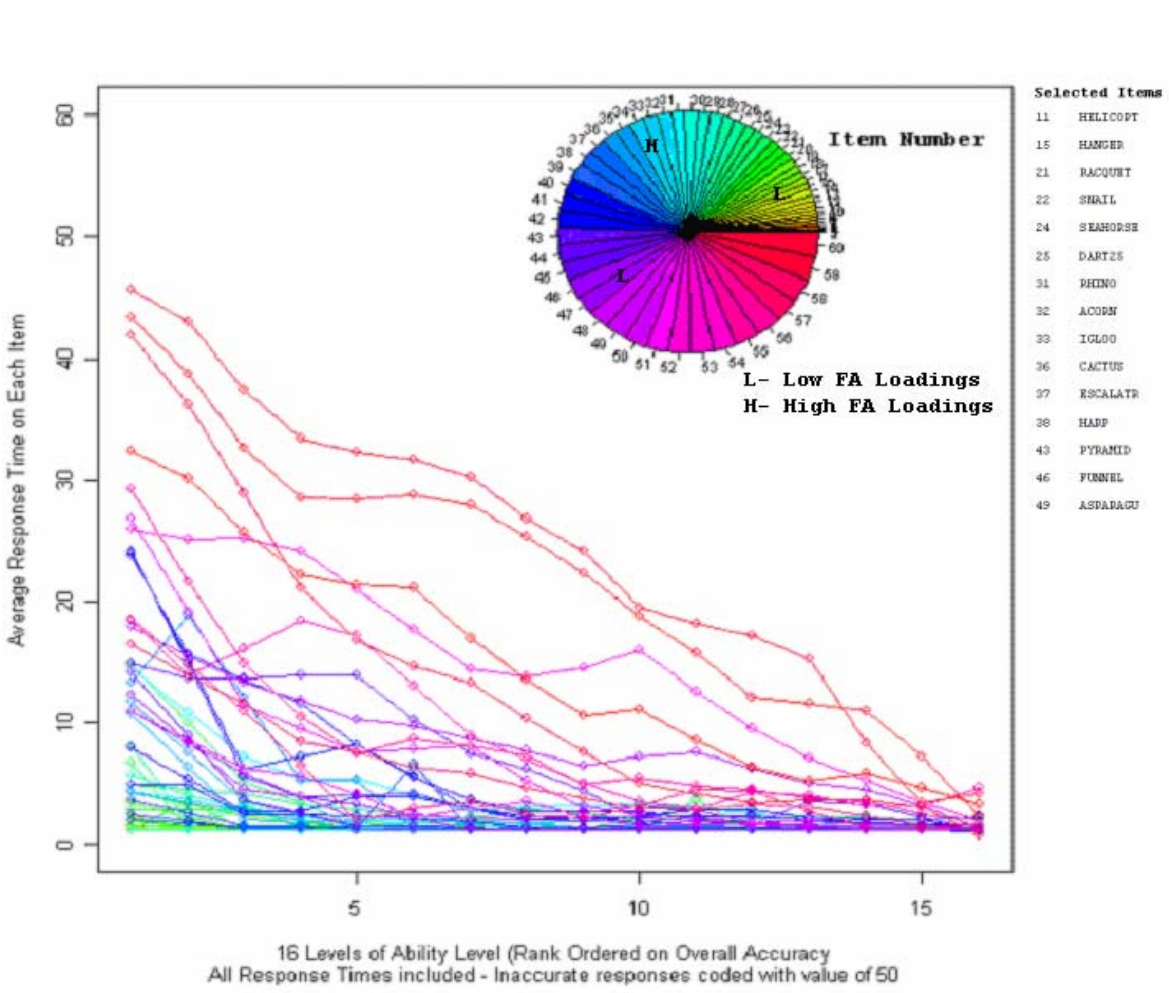
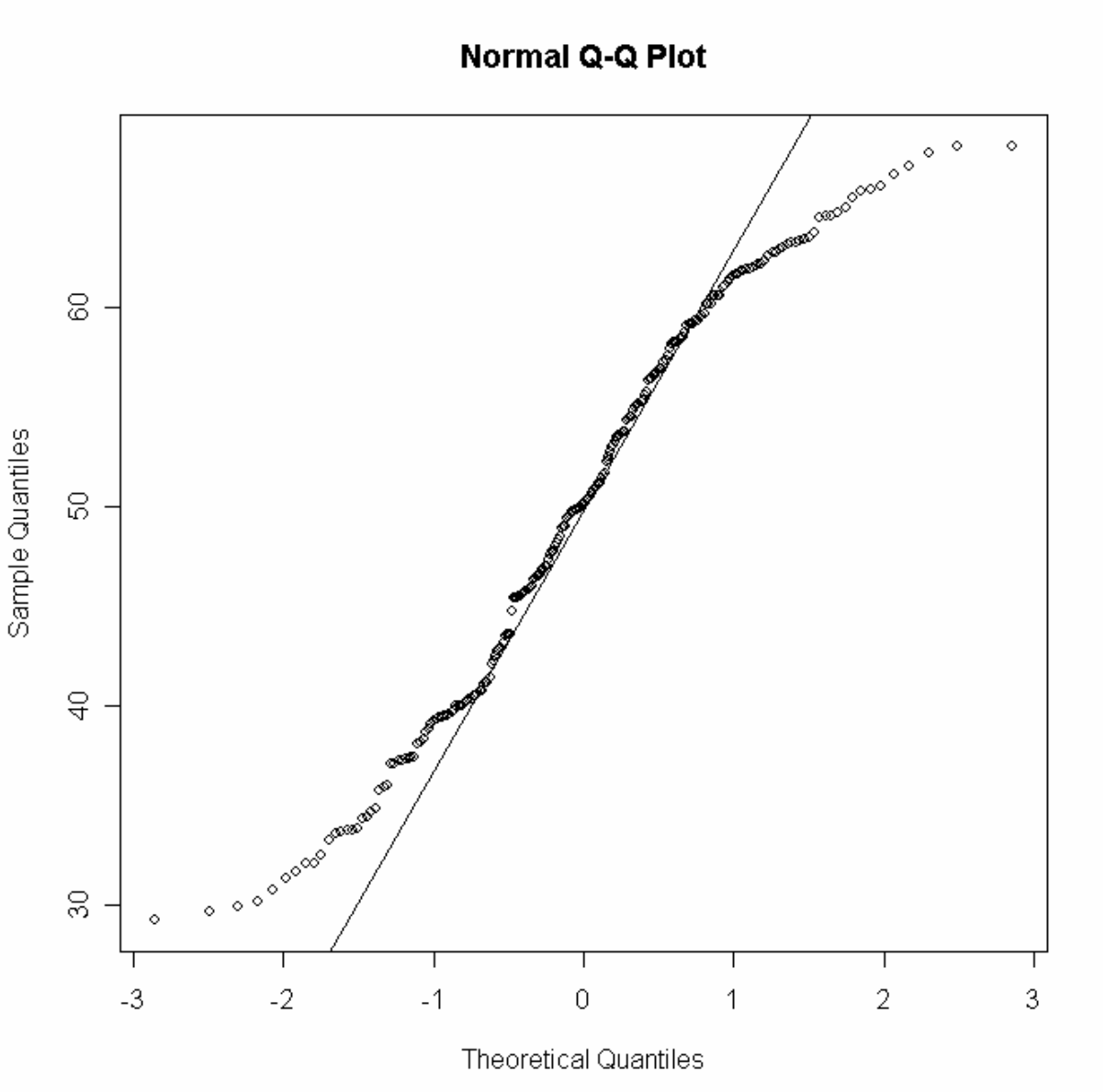


Figure 9. Quantile-quantile plot of the theoretical normal distribution against the empirical quantiles of the adjusted T-scores for the normal comparison group.



APPENDIX A

HEALTH SCREENING WORKSHEET

HEALTH SCREENING WORKSHEET

*Amended Health Screening Exclusion Criteria

1. Stroke or transient ischemic attack
2. History of seizures
3. Parkinson's disease
4. Multiple sclerosis
5. Cerebral palsy
6. Huntington's disease
7. Encephalitis
8. Meningitis
9. Brain surgery
10. Surgery to clear arteries to the brain
11. Diabetes that requires insulin to control
12. Hypertension that is not well controlled
13. Cancer other than skin cancer diagnosed within the past 3 years
14. Shortness of breath while sitting
15. Use of home oxygen
16. Heart attach with changes in memory, ability to talk, or solve problems lasting at least 24 hours afterward.
17. Kidney dialysis
18. Liver disease
19. Hospitalization for mental or emotional problems in past 5 years
20. Current use of medications for mental or emotional problems
21. Alcohol consumption greater than 3 drinks each day
22. Abuse of drugs or medications in past 5 years
23. Treatment for alcohol abuse in past 5 years

24. Unconsciousness for more than one hour other than during surgery
25. Overnight hospitalization because of a head injury
26. Illness causing a permanent decrease in memory or other mental functions
27. Trouble with vision that prevents reading ordinary print even with glasses on
28. Difficulty understanding conversations because of hearing even if wearing a hearing aid
29. Inability to write own name
30. *A diagnosed learning disability
31. *English is not the native *and* primary language
32. *Depends on others for activities of daily living (e.g. does not handle own finances) or lives in a nursing home or institution (e.g. jail)

*The first 29 questions are taken directly from Christensen, Multhaup, Nordstrom, and Voss' (1991) *Health Screening Exclusion Criteria* and the remaining criteria were added by the author.

APPENDIX B

WRITTEN INSTRUCTIONS AND FOR BNT-L ADMINISTRATION AND SCORING

ADMINISTRATION AND SCORING INSTRUCTIONS for BNT-L

Latencies are to be measured in whole seconds using a hand-held digital stopwatch. Response timing begins at the termination of flipping over the stimulus card and ends with the initiation of an accurate verbal response from the participant. The pictures are presented in order beginning with Item 1. No discontinue rule is to be applied. Consecutive errors are permitted to allow recovery of potential word-finding failure. All items are to be administered and scored according to the method set forth.

Participants are presented with a stimulus picture and asked, "Please tell me the most common name for these objects in a single word as fast and accurate as you can. It is important that you try to be both quick AND accurate in your responses."

Responses are recorded up to 50 seconds for each item. The first 20 seconds are administered similar to the standard BNT administration procedures with the response times placed on a timeline with a circle and a slash marked the point on the time line with any appropriate codes (as listed on the BNT-L Scoring Sheet). A semantic prompt, or "stimulus cue," and a "S" mark on the protocol timeline, is to be made if the individual misperceives the item as representing something else (e.g. "umbrella" for mushroom, by offering "it is something to eat") or if it is apparent that the individual lacks recognition of the picture (e.g. "I don't know what that is"). The stimulus cues from the traditional BNT Booklet are printed in brackets under each item on the protocol sheet.

Phonemic cues will be given after 30 seconds and "P" will be placed on appropriate location on the timeline. If the participant correctly names the picture following a phonemic cue, inquiry about a TOT will be asked, "Was this word on the tip of your tongue?" and "Y" or "N"

will be coded accordingly. Phonemic cues will be the underlined portion of the target word on the BNT-L Scoring Sheet.

Subjects may offer unlimited responses within the 50-second time limit. If more than one naming response is made, the final one is scored unless they respond, "It's a ____ or a _____." Then the examiner will ask them to pick one response. All responses will be recorded verbatim in the response booklet. Incorrect responses will be coded as "RN" for Related Name when a person responds with a name similar to the test item (e.g., boat for canoe) and the examiner will say, "*No. We are looking for a better word. Try again.*" "DK" will be logged on the time line for responses of "don't know." Inquiry will be made to determine if the missed item is due to not knowing the word or to linguistic failure. A "V" will be coded if the participant used "verbalizations" as a strategy to find the correct word. After a total of 50 seconds, the examiner will say, "*Let's move on*" and proceed to the next item.

Response codes:

S = Stimulus cue DK = Don't know

P = Phonetic cue RN = Related name

V = Verbalizations SS = Similar Sound

Protocols are scored by adding up the full second response times for all 15 items. Items with a "Don't Know" response are coded with a maximum of 50.

REFERENCES

Albert, M.S., Heller, H.S., & Milberg, W. (1988). Changes in naming ability with age. *Psychology and Aging, 3*, 2 173-178.

Amrhein, P.C. (1995). Evidence for task specificity in age-related slowing: A review of speeded picture-word processing studies. In P.A. Allen & TR. Bashore (Eds.), *Age differences in word and language processing* (pp. 143-170). Amsterdam: Elsevier.

Au, R., Joung, P., Nicholas, M., Kass, R., Obler, L.K., & Albert, M.L. (1995). Naming ability across the lifespan. *Aging and Cognition, 2*(4), 300-311.

Azrin, R.L., Mercury, M.G., Millsaps, C., Goldstein, D., Trejo, T., & Pliskin, N.H. (1996). [Abstract]. Cautionary note on the Boston Naming Test: Cultural considerations. *Archives of Clinical Neuropsychology, 11*(5), 365-366.

Barry, C., Morrison, C.M., & Ellis, A.W. (1997). Naming the Snodgrass Vanderwart Pictures: Effects of age of acquisition, frequency, and name agreement. *Quarterly Journal of Experimental Psychology, 50A*(3), 560-585.

Borod, J.C., Goodglass, H., & Kaplan, E. (1980). Normative data on the Boston Diagnostic Aphasia Examination, Parietal Lobe Battery, and the Boston Naming Test. *Journal of Clinical Neuropsychology, 2*, 209-215.

Bowler, R.M., Thaler, C.D., Law, D., & Becker, C.E. (1990). Comparison of the NEW and CNSB neuropsychological screening batteries. *Neurotoxicology, 11*, 451-464.

Brookshire, R.H. (1971). Effects of trial time and inter-trial interval on naming by aphasic subjects. *Journal of Communication Disorders, 3*, 289-301.

Brookshire, R.H. (1997). *Introduction to neurogenic communication disorders* (5th ed.). St. Louis: Mosby.

Brown, A. S. (1991). A review of the tip-of-the tongue experience. *Psychological Bulletin, 109*(2), 204-223.

Brown, A.S., & Nix, L.A. (1996). Age-related changes in the tip-of-the-tongue experience. *American Journal of Psychology, 109*(1), 79-91,

Brown, R., & McMneill, D., 1966. The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behaviors, 5*, 325-327.

Burke, D.M., & MacKay, D.G. (1997). Memory, language and ageing. *Philosophical Transactions of the Royal Society: Biological Sciences, 353*, 1845-1856.

Burke, D.M., MacKay, D.G., & James, L.E. (2000). Theoretical approaches to language and aging. In T.J. Perfect & E.A. Maylor (Eds.) *Models of cognitive aging*. Oxford: University Press.

Burke, D.M., MacKay, D.G., Worthley, J.S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30*, 237-246.

Calero, M.D., Arnedo, M.L., Navarro, E., Ruiz-Pedrosa, M., & Carnero, C. (2002). Usefulness of a 15-item version of the Boston Naming Test in neuropsychological assessment of low-educational elders with dementia. *Journal of Gerontology, 57b*(2), 187-191.

Christensen, K.J., Multhaup, K.S., Nordstrom, S., & Voss, K. (1991). A cognitive battery for dementia: Development and measurement characteristics. *Psychological Assessment, 3*(2), 168-174.

Cruice, M.N., Worrall, L.E., & Hickson, L.M.H. (2000). Boston Naming Test results for healthy older Australians: A longitudinal and cross-sectional study. *Aphasiology, 14*(2), 143-155.

Deary, I.J., & Der, G. (2005). Reaction time explains IQ's association with death. *American Psychological Society, 16*(1), 64-69.

Dunn, N.D., Russell, S.S., & Drummond, S.S. (1989). Effect of stimulus context and response coding variables on word retrieval performances in dysphasia. *Journal of Communication Disorders, 22*, 209-223.

Efron B., Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ellis, A.W., Lum, C., & Lambon Ralph, M.A. (1996). On the use of regression techniques for the analysis of single case aphasic data. *Journal of Neurolinguistics, 9*, 165-174.

Farmer, A. (1990). Performance of normal males on the Boston Naming Test and the Word Test. *Aphasiology, 4*(3), 293-296.

Fastenau, P.S. (1998). Validity of regression-based norms: An empirical test of the comprehensive norms with older adults. *Journal of Clinical and Experimental Neuropsychology, 20*(6), 906-916.

Fastenau, P.S., Denburg, N.L., & Mauer, B.A. (1998). Parallel short forms of the Boston Naming Test: Psychometric properties and norms for older adults. *Journal of Clinical and Experimental Neuropsychology, 20*(6), 828-834.

Ferman, T.J., Ivnick, R.J., & Lucas, J.A. (1998). Boston Naming Test discontinuation rule: Rigorous versus lenient interpretations. *Assessment, 5*, 13-18.

Feyereisen, P. (1997). A meta-analytic procedure shows an age-related decline in picture naming: Comments on Goulet, Ska, and Kahn (1994). *Journal of Speech, Language, and Hearing Research, 40*(6), 1328-1333.

Feyereisen, P., Demaeght, N., & Samson, D. (1998). Why do picture naming latencies increase with age: General slowing, greater sensitivity to interference, or task-specific deficits? *Experimental Aging Research, 24*, 21-47.

Felmingham, K.L., Baguley, I.J., & Green, A.M. (2004). Effects of diffuse axonal injury on speed of information processing following severe traumatic brain injury. *Neuropsychology, 18*(3), 564-571.

Geary, D.C. (1989). A model for representing gender differences in the pattern of cognitive abilities. *American Psychologist, 44,* 1155-1156.

Georgieff, N., Dominey, P.F., Michel, F., Marie-Cardine, M., & Kalery, J. (1998). Anomia in major depressive state. *Psychiatry Research, 77,* 197-208.

German, D.J. (1991). *Test of word-finding in discourse: Administration, scoring, interpretation, and technical manual*. Allen, TX: DLM.

Geschwind, N., (1967). The varieties of naming errors. *Cortex, 3*, 97-112.

Ginsberg, J.P. (2004). Book and tests reviews: Wechsler Test of Adult Reading. *Applied Neuropsychology, 10*(3), 182-190.

Goldstein, D., Mercury, M., Azin, R., et al. (2000). *Cautionary note on the Boston Naming Test: Cultural considerations*. Paper presented at the 28th annual meeting of the International Neuropsychological Society, Denver, CO.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *The assessment of aphasia and related disorders* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Goodglass, H., & Wingfield, A. (1997). Word-finding deficits in aphasia: Brain-behavior relations and clinical symptomatology. In H. Goodglass & A. Wingfield (Eds.), *Anomia: Neuroanatomical and cognitive correlates* (pp. 5-27). San Diego, CA: Academic Press.

Goodglass, H., Wingfield, A., & Hyped, M.R. (1998). The Boston corpus of aphasic naming errors. *Brain and Language, 64,* 1-27.

Gordon, (1997). In H. Goodglass & A. Wingfield (Eds.), *Anomia: Neuroanatomical and cognitive correlates* (pp. XX). San Diego, CA: Academic Press.

Goulet, P., Ska, B., & Kahn, H.J. (1994). Is there a decline in picture naming with advancing age? *Journal of Speech and Hearing Research, 37,* 629-644.

Hamby, S.L., Bardi, C.A., & Wilkins, J.W. (1997). Neuropsychological assessment of relatively intact individuals: Psychometric lessons from an HIV+ sample. *Archives of Clinical Neuropsychology, 12*(6), 545-556.

Harrell, F.E., Lee, K.L., & Mark, D.B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15,* 361-387.

Hawkins, K.A., & Bender, S. (2002). Norms and the relationship of Boston Naming Test performance to vocabulary and education: A review. *Aphasiology, 16*(12), 1143-1153.

Hawkins, K.A., Sledge, W.H., Orleans, J.E., Quinlan, D.M, Rakfeldt, J., & Huffman, R.E. (1993). Normative implications of the relationship between reading vocabulary and Boston Naming Test performance. *Archives of Clinical Neuropsychology, 8*, 525-537.

Heaton, R.K., Avitable, N., Grant, I.,, & Matthews, C.G. (1999). Further cross validation of regression-based neuropsychological norms with an update for the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology, 21*(4), 572-582.

Heaton, R., Grant, I., & Matthews, C. (1991). *Comprehensive norms for an expanded Halstead-Reitan Neuropsychological Battery: Demographic corrections, research findings and clinical applications*. Odessa, FL: Psychological Assessment Resources.

Henderson, L.W., Frank, E., Pigatt, T;, Abramson, R.K., & Houston, M. (1998). Race, gender, and educational level effects on Boston Naming Test scores. *Aphasiology, 12*(10), 901-911.

Hickman, S.E., Howieson, D.B., Dame, A., Sexton, G., & Kaye, J. (2000). Longitudinal analysis of the effects of the aging process on neuropsychological test performance in the healthy young-old and oldest-old. *Developmental Neuropsychology, 17*(3), 323-337.

Hodgson, C., & Ellis, A.W. (1998). Last in, first to go: Age of acquisition and naming in the elderly. *Brain and Language, 64*, 146-163.

Hooper, H. E. (1983). *The Hooper Visual Organization Test manual*. Los Angeles: Western Psychological Services.

Hubley, A., & Tombaugh, T.N. (1998). Norms for the 60-item Boston Naming Test (BNT) for cognitively intact individuals aged 25-90 years. [Abstract]. *Archives of Clinical Neuropsychology, 13*(1), 101.

Indefrey, P., & Levelt, W.J.M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*, 101-144.

James, L., & Burke, D.M. (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*(6), 1378-1391.

James, W. (1893). *The principles of psychology: Vol. I*. New York: Henry Holt.

Kaplan, E., Goodglass, H., & Weintraub, S. (1978). The Boston Naming Test, Experimental Edition. Boston: Kaplan and Goodlass.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). The Boston Naming Test. Philadelphia: Lea and Febiger.

Kemeny, S., Xu, Jian, Park, G.H., Hosey, L.A., Wettig, C.M., & Braun, A.R. (2006). Temporal dissociation of early lexical access and articulation using a delayed naming task – An fMRI study. *Cerebral Cortex, 16*(4), 587-595.

Kent, P.S., & Luszcz, M.A. (2002). A review of the Boston Naming Test and multiple-occasion normative data for older adults on 15-item versions. *Clinical Neuropsychologist, 16*(4), 555-574.

Killgore, W.D., & Adams, R.L. (1998). Vocabulary ability and the Boston Naming Test performance: Interpretative guidelines. [Abstract]. *Archives of Clinical Neuropsychology, 13*(1), 31.

Kim, H., & Na, D.L. (1999). Normative data on the Korean version of the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology, 21*(1), 127-133.

Kohn, S., & Goodglass, H. (1985). Picture-naming in aphasia. *Brain and Language, 24*, 266-283.

LaBarge, E., Edwards, D., & Knesevich, J.W. (1986). Performance of normal elderly on the Boston Naming Test. *Brain and Language, 27*, 380-384.

Lambon Ralphe, M.A., Moriarty, L., & Sage, K. (2002). Anomia is simply a reflection of semantic and phonological impairments: Evidence from a case-series study. *Aphasiology, 16*(1/2), 56-82.

Lansing, A.E., Ivnick, R.J., Cullum, C.M., & Randolph, C. (1999). An empirically derived short form of the Boston Naming Test. *Archives of Clinical Neuropsychology, 1*, 481-487.

Larver, G.D., & Burke, D.M. (1993). Why do semantic priming effects increase in old age? A meta-analysis. *Psychology and Aging, 8*, 34-43.

Le Dorze, G., & Durocher, J. (1992). The effects of age, educational level, and stimulus length on naming in normal subjects. *Journal of Speech and Language Pathology and Audiology, 16*, 21-29.

Lesk, V.E., & Womble, S.E. (2004). Caffeine, priming, and tip of the tongue: Evidence for plasticity of the phonological system. *Behavioral Neuroscience, 118*(3), 453-461.

Levelt, W.J.M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National. Academy of Sciences, 98*(23), 13464-13471.

Lezak, M.D. (1991). Emotional impact of cognitive inefficiencies in mild head trauma [Abstract]. *Journal of Clinical and Experimental Neuropsychology, 13*, 23.

Lezak, M.D. (2004). *Neuropsychological assessment* (4th ed.). Oxford: Oxford University Press.

Lopez, M.N., Arias, G.P., Hunter, M.A., Charter, R.A., & Scott, R.R. (2003). Boston Naming Test: Problems with administration and scoring. *Psychological Reports, 92*, 468-472.

Loring, D.W. (1999). *INS dictionary of neuropsychology*. New York: Oxford University Press.

Lovelace, E.A., & Twohig, P.T. (1990). Healthy older adults' perception of their memory functioning and the use of mnemonics. *Bulletin of the Psychonomic Society, 28*(2), 115-118.

Luce, D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford: Oxford University Press.

Mack, W.J., Freed, D.M., Williams, B.W., & Henderson, V.W. (1992). Boston Naming Test: Shortened versions for use in Alzheimer's disease. *Journal of Gerontology: Psychological Sciences, 47,* 154-158.

MacKay, A., Connor, L.T., Albert, M.L., & Obler, L.K. (2002). Noun and verb retrieval in healthy aging. *Journal of the International Neuropsychological Society, 8*, 764-770.

MacKay, A., Connor, L.T., & Storandt, M. (in press). Dementia does not explain correlation between age and scores on the Boston Naming Test. *Archives of Clinical Neuropsychology*.

Madden, D.J. (1988). Adult age differences in the effects of sentence context and stimulus degradation during visual word recognition. *Psychology and Aging, 3,* 167-172.

Martin, M., & Zimprich, D. (2003). Are changes in cognitive functioning in older adults related to changes in subjective complaints? *Experimental Aging Research, 29,* 335-352.

McDonald, R.P. (1999). *Unified test theory*. New Jersey: Earlbaum Associates.

McGurn, B., Starr, J.M., & Topfer, J.A. (2004). Pronunciation of irregular words is preserved in dementia: Validating premorbid IQ estimation. *Neurology, 62,* 1184-1186.

Mehta, K.M., Yaffe, K., Covinsky, K.E. (2002). Cognitive impairment, depressive symptoms, and functional decline in older people. *Journal of American Geriatric Society, 50*, 1045-1050.

Mitrushina, M.N., Boone, K., B., D'Elia, L.F. (1999). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.

Mitchell, D.B. (1989). How many memory systems? Evidence from aging. *Journal of Experimental Psychology: Learning , Memory, and Cognition, 15,* 31-49.

Morrison, C., M., Chappell, T.D., & Ellis, A.W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology, 50A*(3), 528-559.

Morrison, C.M., Ellis, A.W., & Quinlan, P.T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory and Cognition, 20,* 705-714,

Nicolas, M., Barth, C., Obler, L.K., Au, R., & Albert, M.L. (1997). Naming in normal aging and dementia of the Alzheimer's type. In H. Goodglass & A. Wingfield (Eds.) *Anomia: Neuroanatomical and cognitive correlates* (pp. 166-188). New York: Academic Press.

Nicholas, M., Brookshire, R.H., MacLennan, D.L., Schumacher, J.G., & Porrazzo, S.A. (1989). Revised administration and scoring procedures for the Boston Naming Test and norms for non-brain-damaged adults. *Aphasiology, 3*(6), 569-580.

Nicholas, M., Obler,, L., Albert, M.L., & Goodglass, H. (1985). Lexical retrieval in healthy aging. *Cortex, 21,* 595-606.

Nyberg, L., Backman, L., Erngrund, K., Olofsson, U., & Nilsson, L. (1996). Age differences in episodic memory, semantic memory, and priming: Relationships to demographic, intellectual, and biological factors. *Journal of Gerontology, 51B,* 234-240.

Obler, L.K., & Albert, M.L. (1979). Action Naming Test, (Experimental Edition). Boston: VA Medical Center.

Obler, L.K., & Albert, M.L. (1985). Language skills across adulthood. In J. Birren & K.W. Schaie (Eds.), *The psychology of aging*. New York: Van Nostrand Reinhold.

Orange, J.B., & Purves, B. (1996). Conversational discourse and cognitive impairment: Implications for Alzheimer's disease. *Journal of Speech Language Pathology and Audiology, 20,* 139-151.

Perbal, S., Couillet, J., Azouvi, P, & Pouthas. (2003). Relationships between time estimation, memory, attention, and processing speed in patients with severe traumatic brain injury. *Neuropsychologia, 41,* 1599-1610.

Pontón, M.O., Satz, P., Herrera, L. et al. (1996). Normative data stratified by age and education for the Neuropsychological Screening Battery for Hispanics (NeSBHIS): Initial report. *Journal of the International Neuropsychological Society, 2,* 96-104.

Ponds, R., van Boxtel, M., & Jolles, J. (2000). Age-related changes in subjective cognition functioning. *Educational Gerontology, 26,* 67-81.

Poon, L.W., & Fozard, J.L., (1978). Speed of retrieval from long-term memory in relation to age, familiarity, and datedness of information. *Journal of Gerontology, 33,* 711-71.

Randolph, C., Lansing, A.,E., Ivnick, R.J., Cullum, C.M., & Hermann, B.P. (1999). Determinants of confrontation naming performance. *Archives of Clinical Neuropsychology, 14*(6), 489-496.

Reason, J. T. (1984). Lapses of attention in everyday life. In R. Parasuraman & D.R, Davies (Eds.), *Varieties of attention* (pp. 515-549). San Diego, DA: Academic Press.

Rees, L.M., Tombaugh, T.N., & Boulay, L. (2001). Depression and the Test of Memory Malingering (TOMM). *Archives of Clinical Neuropsychology, 16,* 501-506.

Roberts, P.M., Garcia, L.J., Desrochers, A., & Hernandez. (2002). English performance of proficient bilingual adults on the Boston Naming Test. *Aphasiology, 16*, 635-645.

Ross, T.P., & Lichtenberg, P.A. (1997). Expanded normative data for the Boston Naming Test in an urban medical sample of elderly adults [Abstract]. *Journal of the International Neuropsychological Society, 3,* 70.

Russell, E.W., & Starkey, R.I. (1993). Halstead Russell Neuropsychological Evaluation System (HRNES). Los Angeles: Western Psychological Services.

Salthouse, T.A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403-428.

Schmitter-Edgecombe, M., Vesneski, M., & Jones, D.W.R. (2000). Aging and word-finding: A comparison of spontaneous and constrained naming tests. *Archives of Clinical Neuropsychology, 58*, 397-405.

Smith, G., & Rush, B.K. (2006). Normal aging and mild cognitive impairment. In D.K. Attix and K.A. Welsh-Bohmer (Eds.) *Geriatric neuropsychology: Assessment and intervention* (pp. 5-55). New York: Guilford Press.

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 174-215.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms and commentary* (2nd ed.). New York: Oxford University Press.

Stern, C., Prather, P., Swinney, D., & Zurif, E. (1991). The time course of automatic lexical access and aging. *Brain and Language, 40,* 359-372.

Sunderland, A., Watts, K., Baddeley, A.D., & Harris, J.E. (1986). Subjective memory assessment and test performance in elderly adults. *Journal of Gerontology, 41*(3), 376-384.

Tartter, V.C. (1998). *Language processing in atypical populations*. London: Sage Publications.

Thomas, J.C., Fozard, J.L., & Waugh, N.C. (1977). Age-related differences in naming latency. *American Journal of Psychology, 90*(3), 499-509.

Thompson, L.L., & Heaton, R.K. (1989). Comparison of different versions of the Boston Naming Test. *Clinical Neuropsychologist, 3*(2), 194-192.

Tombaugh, T. N. (1996). Test of Memory Malingering (TOMM). Multi-Health Systems. Toronto, Ontario.

Tombaugh, T.N., & Hubley, A.M. (1997). The 60-item Boston Naming Test: Norms for cognitively intact adults 25-88 years. *Journal of Clinical and Experimental Neuropsychology, 19*, 922-932.

Tsang, H.L., & Lee, T. M.C. (2003). The effect of ageing on confrontational naming ability. *Archives of Clinical Neuropsychology, 18*, 81-89.

Tsolaki, M., Tsantali, E., Lekka, S., Kiosseoglou, G., & Kazis, A. (2003). Can the Boston Naming Test be used as clinical tool for differential diagnosis in dementia? *Brain and Language, 87*, 185-186.

U.S. Census Bureau; Census 2000, Summary File 3 (SF 3); Using American Factfinder. Retrieved 2 February 2006 from http://factfinder.census.gov

Van Gorp, W.G., Satz, P., Kiersch, M.E., & Henry, R. (1986). Normative data on the Boston Naming Test for a group of normal older adults. *Journal of Clinical and Experimental Neuropsychology, 8*(6), 702-705.

Vigneau, F., Blanchet, L., Loranger, M., & Pepin, M. (2002). Response latencies measured on IQ tests: Dimensionality of speed indices and the relationship between speed and level. *Personality and Individual Differences, 33*, 165-182.

Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology, 48A*, 822-848.

Watson, M.E., Welsh-Bohmer, K.A., Hoffman, J.M., Lowe, V., & Rubin, D.C. (1999). The neural basis of naming impairments in Alzheimer's disease revealed through positron emission tomography. *Archives of Clinical Neuropsychology, 14*(4), 347-357.

Wechsler Test of Adult Reading. (2003). San Antonio: The Psychological Corporation, Harcourt Assessment Company.

Welch, L.W., Doneau, D., Johnson, S., & King, D. (1996). Educational and gender normative data for the Boston Naming Test in a group of older adults. *Brain and Language, 53*, 260-266.

Whitfield, K.E., Fillenbaum, G.G., Pieper, C., Albert, M.S., Berkman, L. F., Blazer, D.G., Rowe, J.W., & Seeman, T. (2000). The effects of race and health-related factors on naming and memory. *Journal of Aging and Health, 12*(1), 69-89.

Wilcox, R.R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314.

Williams, B.W., Mack, W., & Henderson, V.W. (1989). Boston Naming Test in Alzheimer's disease. *Neuropsychologia, 27*(8), 1073-1079.

Worrall, L.E., Yiu, M.L., Hickson, L.M.H., & Barnett, H.M. (1995). Normative data for the Boston Naming Test for Australian elderly. *Aphasiology, 9*, 541-551.

Yaniv, I., & Meyer, D.E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 187-205.